DOCUMENT RESUME

ED 060 917 52 LI 003 608

AUTHOR Mignon, Edmond; Travis, Irene
TITLE LABSEARCH: ILR Associative Search System Terminal Users' Manual. Final Report.
INSTITUTION California Univ., Berkeley. Inst. of Library Research.
SPONS AGENCY Office of Education (DHEW), Washington, D.C. Bureau of Research.
BUREAU NO BR-7-1085
PUB DATE Sep 71
GRANT OEG-1-7-071085-4286
NOTE 87p.; (5 References)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Automation; Data Bases; Electronic Data Processing; Indexing; *Information Processing; *Information Retrieval; *Library Education; *Library Science; Manuals; Research; Search Strategies
IDENTIFIERS *University of California Berkeley

ABSTRACT

The results of the second 18 months (December 15, 1968-June 30, 1970) of effort toward developing an Information Processing Laboratory for research and education in library science is reported in six volumes. This volume contains: basic operating instructions, commands, scoring measures of association and a subject authority list. The data base consists of journal articles in the field of information science. Indexing was done using a controlled vocabulary from the subject authority list. The two kinds of automatic searching procedures for the identification and retrieval of documents described are: direct match and associative search. (Other volumes of this report are available as LI 003607 and LI 003609 through 003611). (Author/NH)

ED 060917

WI
OE-BR-7-1085

FINAL REPORT
Project No. 7-1085
Grant No. OEG-1-7-071085-4286

LABSEARCH:
ILR ASSOCIATIVE SEARCH SYSTEM
TERMINAL USERS' MANUAL

By

Edmond Mignon
Irene Travis

Institute of Library Research
University of California
Berkeley, California 94720

September 1971

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

LI 003 608

ED 060917

U.S. DEPARTMENT OF HEALTH, EDUCATION, & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

PA-52
BR-7-1085

FINAL REPORT.
Project No. 7-1085
Grant No. OEG-1-7-071085-4286

LABSEARCH:
ILR ASSOCIATIVE SEARCH SYSTEM
TERMINAL USERS' MANUAL

By

Edmond Mignon
Irene Travis

Institute of Library Research
University of California
Berkeley, California 94720

September 1971

The research reported herein was performed pursuant to a grant with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

LI 003 608

89p.

TABLE OF CONTENTS

		Page
1.	INTRODUCTION	1
	1.1 Overview	1
	1.2 The Corpus	1
	1.3 The Indexing	2
	1.4 The Accession Numbers	2
	1.5 The Search Modes	2
	1.6 The Association Files	2
	1.7 The Search Request	4
	1.7.1 The Boolean Expression	4
	1.7.2 Weights	5
	1.8 Output	5
	1.9 Further Information	6
2.	BASIC OPERATING INSTRUCTIONS	7
	2.1 The Log-in Procedure	7
	2.2 The Normal Pass	7
	2.3 The Six Questions Individually Discussed	7
3.	COMMANDS	13
	3.1 Introduction	13
	3.2 GO TO--Commands for Changing the Normal Program Flow	13
	3.3 EDIT and SHOW--Commands for Modifying and Displaying the Boolean Request Expression and the Other Request Specifications	17
	3.3.1 EDIT	17
	3.3.2 SHOW	18
	3.4 DISPLAY DOCUMENTS, SORTS, SORTD--Commands for Sorting, Selecting, and Displaying the Set of Retrieved Document Accession Numbers and Relevance Scores	18
	3.4.1 DISPLAY DOCUMENTS	18
	3.4.2 SORTA and SORTD	19
	3.5 GET, DISPLAY, and RETRIEVE--Commands for Displaying Data Files	20
	3.5.1 GET	20
	3.5.2 DISPLAY	21
	3.5.3 RETRIEVE	22
4.	SCORING	25
	4.1 General	25
	4.2 Requests with AND Operator	25
	4.3 Requests with OR Operator	27

TABLE OF CONTENTS (Cont.)

Page

4.5.1 The Effect of Weights on Relevance Scores..... 29
4.5.2 The Effects of Term Weights on the Ranking of the Retrieved Set..... 30

5. ASSOCIATION FILES 31

5.1 Preliminaries..... 31
5.2 Notation..... 31
5.3 Excess Over Independence Value..... 32
5.4 Association Measures..... 34
5.5 Notes on Individual Measures..... 34
5.5.1 KUHNSY..... 36
5.5.2 KUHNSW..... 38
5.5.3 KUHNSS..... 42
5.5.4 KUHNSG and KUHNSL..... 43
5.5.5 The Normalized Measures, KUHNSGN and KUHNSSN..... 43
5.5.6 DOYLE..... 45

INDEX OF KEY DEFINITIONS 49

APPENDICES

Page

1. LABSRC3C SUBJECT AUTHORITY LIST 51

2. LIST OF INDEX TERM FREQUENCIES IN THE CORPUS 59

3. INDEX TERM FREQUENCY LIST SORTED ON FREQUENCY OF REFERENCE 67

4. A CLASSIFICATION OF THE SUBJECT AUTHORITY LIST 75

LIST OF FIGURES

Figure	Title	Page
SECTION 2: BASIC OPERATING INSTRUCTIONS		
1.	Example of a LABSRC3C Output Display	11
SECTION 3: COMMANDS		
2.	MASTERI Record for Document B1218	20
3.	Typical Association Table	21
4.	MASTERA Record for Document B1218	24
SECTION 5: ASSOCIATION FILES		
5.	DOYLE Coefficient of .2500 for n_1 = 20 as a Function of min(n_1, n_2)	46

LIST OF TABLES

Table	Title	Page
SECTION 1: INTRODUCTION		
1.	Comparison of the Interpretation of Boolean Operators in Direct Match Mode and Associative Retrieval Mode . .	3
SECTION 3: COMMANDS		
2.	Index to the Commands	14
3.	Default Options	16
SECTION 5: ASSOCIATION FILES		
4.	Relationship Between Expressions A and B	32
5.	Reference Chart of Kuhns Measures in LABSRC3C	35
6.	Summarized Comparison of Doyle and Kuhns Measures . . .	47

FOREWORD

This report contains the results of the second 18 months (December 15, 1968 - June 30, 1970) of effort toward developing an Information Processing Laboratory for research and education in library science. The work was supported by a grant (OEG-1-7-071085-4286) from the Bureau of Research of the Office of Education, U.S. Department of Health, Education, and Welfare and also by the University of California. The principal investigator was M.E. Maron, Professor of Librarianship.

This report is being issued as six separate volumes by the Institute of Library Research, University of California, Berkeley. They are:

- Maron, M.E. and Don Sherman, et al. An Information Processing Laboratory for Education and Research in Library Science: Phase 2.

  Contents--Introduction and Overview; Problems of Library Science; Facility Development; Operational Experience.

- Mignon, Edmond and Irene L. Travis. LABSEARCH: ILR Associative Search System Terminal Users' Manual.

  Contents--Basic Operating Instructions; Commands; Scoring Measures of Association; Subject Authority List.

- Meredith, Joseph C. Reference Search System (REFSEARCH) Users' Manual.

  Contents--Rationale and Description; Definitions; Index and Coding Key; Retrieval Procedures; Examples.

- Silver, Stephen S. and Joseph C. Meredith. DISCUS Interactive System Users' Manual.

  Contents--Basic On-Line Interchange; DISCUS Operations; Programming in DISCUS; Concise DISCUS Specifications; System Author Mode; Exercises.

- Smith, Stephen F. and William Harrelson. TMS: A Terminal Monitor System for Information Processing.

  Contents--Part I: Users' Guide - A Guide to Writing Programs for TMS
  Part II: Internals Guide - A Program Logic Manual for the Terminal Monitor System

- Aiyer, Arjun K. The CIMARON System: Modular Programs for the Organization and Search of Large Files.

  Contents--Data Base Selection; Entering Search Requests; Search Results; Record Retrieval Controls; Data Base Generation.

Because of the joint support provided by the File Organization Project (OEG-1-7-071083-5068) for the development of DISCUS and of TMS, the volumes concerned with these programs are included as part of the final report for both projects. Also, the CIMARON System, whose development was supported by the File Organization Project, has been incorporated into the Laboratory operation and therefore, in order to provide a balanced view of the total facility obtained, that volume is included as part of this Laboratory project report. (See Shoffner, R.M., et al., The Organization and Search of Bibliographic Records in On-Line Computer Systems: Project Summary.)

ACKNOWLEDGMENTS

Many people from both the School of Librarianship and the Institute of Library Research contributed to the LABSEARCH program.

We would like to thank Mary Wilson and Lois Harzfeld for their help in organizing the manual and selecting illustrative examples. Keith Stirling contributed analytical material to Chapter 5, and the Classified Term List is largely the work of Connie Farrow. Joseph C. Meredith, Allan Humphrey, and Don Sherman gave us astute criticisms of preliminary drafts of the manual and also helped us in countless other ways.

The LABSEARCH program was created by C.V. Ravi, who spent much time and effort assisting us and other initial users of the program. Our greatest debt is to the students of the School of Librarianship, whose spirited and candid observations on their experience with LABSEARCH were our most valuable guide in determining the contents and character of this manual.

Finally, we wish to thank Ellen Drapkin, Carole Fender, Bettye Geer, Linda Herold, Jan Kumataka, and Rhozalyn Perkins for their work in the preparation of these pages.

1. INTRODUCTION

1.1 Overview

LABSRC3C is a search program to teach and demonstrate automatic retrieval techniques. In response to requests in the form of Boolean expressions, LABSRC3C carries out a search which may be automatically expanded using terms associated with the original request terms. The degree of association of one term with another is determined by using one of eight statistical association measures operating on index-term co-occurrence data. The program computes a probable relevance score for each document it retrieves. It operates interactively via Sanders video terminals and utilizes a cathode ray tube (CRT) screen for all user and program displays. It offers a range of options for formulating search requests and for modifying search strategy.

The materials for constructing the data base are

1. A set of documents, hereafter called the corpus. The corpus consists of journal articles in the field of information science.

2. A controlled vocabulary used for indexing the documents called the Subject Authority List.

Each journal article is assigned an accession number and indexed. A machine record for each document, consisting of its accession number and the index terms that have been applied to it, is generated. The file of all these index-term records is called the MASTERI file, and it is on this file that the searches are carried out. For an example of a record from the MASTERI file, see Fig. 2 on page 20.

An auxiliary file, called MASTERA, contains the abstracts of the documents in the corpus and is stored in a separate area of disc storage. The MASTERA file is not used directly for searching, but abstracts may be readily requested on-line by the user and displayed on the Sanders video terminal.

1.2 The Corpus

The records in the MASTERI file describe a set of 400 articles on information science published since 1957, and chosen chiefly from the primary research journals of ACM, ASIS, ASLIB, AFIPS, and (more selectively) other journals and symposia of similar stature and character. At the present time, the most recent articles in the corpus date from 1968, but further additions to this collection are to be made from time to time.

The texts of the documents are not stored in the computer, but the documents themselves, in microfiche format, are kept in the Information Processing Laboratory and are easily accessible. The Laboratory also contains reading apparatus as well as a microfiche

copier.

1.3 The Indexing

Index terms from the Subject Authority List have been manually assigned to documents in the corpus, with an average indexing depth of approximately 15:1. The index-terms are not all single-word expressions. Some of them are pre-coordinated phrases, such as manual indexing or state-of-the-art; thus technically speaking they are descriptors rather than index terms in the conventional sense.

The Subject Authority List is included in Appendix 1.

1.4 The Accession Numbers

Each document is manually assigned a five-character accession number, consisting of a letter followed by four digits. The letters used for these accession numbers are A, B, and X. There is no classificatory significance to these letters.

1.5 The Search Modes

It is possible to specify either of two kinds of automatic searching procedures for the identification and retrieval of documents in response to a search request: direct match or associative search.

When the system is searching in direct match mode, it will retrieve only those documents whose index terms exactly match the specifications of the search request. Although such a search will be generally satisfactory in the sense that most or all of the documents retrieved will have a readily recognizable correspondence with the topics specified by the index terms in the search request, the search may nevertheless overlook some relevant documents because of variations in indexing or search request formulation.

In order to get around this difficulty, LABSRC3C offers a means for extending the search by putting the system into associative retrieval mode. In this procedure the system automatically expands the request and searches not only for documents whose index terms match those of the request, but also for documents indexed under terms which have a high statistical association with the terms in the original request. This automatically elaborates the request in a direction which has high probability of retrieving additional relevant documents. The method is not foolproof since it is based on probability rather than certainty, but the assumption is that the probability of retrieving additional relevant documents will be increased by adding associated terms to those specified in the request.

1.6 The Association Files

Two terms are said to be positively associated if they are used

Table 1

Comparison of the Interpretation of Boolean Operators in Direct Match Mode and Associative Retrieval Mode
('A' and 'B' stand for any two index terms)

Request	Direct Match	Associative
'A' and 'B'	Document must be indexed under <u>both</u> A and B to be retrieved.	Documents must be indexed under one of the three following combinations: (1) both A and B, (2) A and a term highly associated with B, (3) a term highly associated with A and also one highly associated with B.
'A' or 'B'	Document must be indexed under <u>either</u> A or B.	Documents must be indexed under A or B or at least one term which is highly associated with A or B.
'A' and not 'B'	Document must be indexed under A, but not indexed under B.	Documents must be indexed under A or at least one term highly associated with A, but it must not be indexed under B. NOT operators are not expanded in associative mode.

jointly in the indexing of documents more frequently than would be expected by random chance. An association measure is an algebraic formula for calculating and giving a numerical representation of the degree of association between a pair of index terms. The value obtained from this calculation is called an association value. Since the determination of this value depends on occurrences of single terms and co-occurrences of term pairs rather than the meanings of the terms, it is a statistical rather than a semantic measure of the closeness of the terms.

There are many different ways of calculating association, and LABSRC3C permits one to choose from a repertory of association measures, which differ from each other (in some cases quite strikingly) in their properties. The association measures may differ not only in the quantities that they compute for association values, but also in their determination of which terms are found to be most highly associated. Thus different measures may often lead to different retrieval results when searching in associative mode. The individual measures are described and compared in Chapter 5.

For searching in associative mode, LABSRC3C contains a set of association files, one for each association measure. Each file is made up of association tables, one table for each term in the Subject Authority List. Each table consists of an index term, the four other terms most highly associated with it, and their computed association values. The association value of the index term with itself under each formula is also computed and stored; therefore, there are five association values in all in each table. An example of an association table is given and discussed on page 21.

1.7 The Search Request

1.7.1 The Boolean Expression

LABSRC3C provides automatic retrieval of document accession numbers in response to requests submitted in the form of Boolean expressions. The components of a Boolean expression are "Boolean" operators and operands. The operands are, in this case, single index terms from the Subject Authority List or Boolean expressions combining such terms. The three logical or "Boolean" operators are AND, OR, and NOT. Since these operators differ considerably in their effect on the retrieval output, some care must be taken in their application. Compare the following results.*

Ex. 1. 'AUTO. INDEXING' AND 'MANUAL INDEXING'. This request will cause LABSRC3C to retrieve only those documents that are indexed under both of the terms 'AUTOMATIC INDEXING' and 'MANUAL INDEXING'.

Ex. 2. 'AUTO. INDEXING' OR 'MANUAL INDEXING'. In response to this request, the system will retrieve all documents indexed under at least one of these two terms, including documents indexed under both.

---

*The interpretation of these examples assumes searching in Direct Match Mode.

Ex. 3. 'AUTO. INDEXING' AND NOT 'MANUAL INDEXING'. In this case the system will retrieve only those documents indexed under 'AUTO. INDEXING' that are not also indexed under 'MANUAL INDEXING'.

The Boolean expression need not be limited to two index terms. It may be of considerable length and complexity, including parenthetical expressions.

Ex. 4. (('INFO. SCIENCE' OR 'CYBERNETICS') AND 'SOCIAL IMPLIC.') AND NOT ('CURRICULUM' OR 'EDUCATION')

Interpretation: This request will cause the system to retrieve documents that are indexed under both 'INFO. SCIENCE' and 'SOCIAL IMPLIC.' or both 'CYBERNETICS' and 'SOCIAL IMPLICATIONS', provided they are not also indexed under either 'CURRICULUM' or 'EDUCATION'.

The request would be completely changed if the parentheses were moved or removed. For example:

'INFO. SCIENCE' OR ('CYBERNETICS' AND 'SOCIAL IMPLIC.') AND NOT 'CURRICULUM' OR 'EDUCATION'.

Interpretation: This request will retrieve documents indexed under (1) both 'CYBERNETICS' and 'SOCIAL IMPLIC.' or else (2) 'INFO. SCIENCE' or else (3) 'EDUCATION' providing that none of them are indexed under 'CURRICULUM'.

The procedures for entering Boolean expressions are discussed in Chapter 2.

1.7.2 Weights

Weights may be assigned to operands consisting of either single terms or parenthetical expressions or to individual terms within the operands. In general the weights reduce the value which the weighted term contributes to the relevance computation. The complexities of the effects of these weights on retrieval are discussed in Chapter 4. Procedures for inputting weights are discussed in Chapter 2.

1.8 Output

LABSRC3C responds to a request by searching the MASTERI file and producing a list of the accession numbers of the documents that satisfy the specifications of the request. If scoring is requested, the program will compute and display a relevance score for each document retrieved. The relevance score represents the degree to which the indexing of the retrieved document matches the terms of the request. A detailed explanation of how probable relevance scores are calculated is given in Chapter 4.

Unless otherwise specified by the user, the retrieved documents are listed in accession number order. The user, however, may

have the program sort the output by relevance score in either ascending or descending order by using the SORTA and SORTD commands. (See Sec. 3.4.2)

1.9 Further Information

For an amplified discussion of the rationale and workings of LABSRC3C, consult Chapter 5 of M.E. Maron et al., An Information Processing Laboratory for Education and Research in Library Science: Phase I (Berkeley: Institute of Library Research, July 1969). Copies of this report are available in the Library School Library, University of California, Berkeley.

2. BASIC OPERATING INSTRUCTIONS

2.1 The Log-in Procedure

Since the log-in procedure is changed from time to time, it does not seem worthwhile to include it in this manual. Current instructions are usually available as hand-outs in the lab, and the lab supervisor can give on-the-spot help.

2.2 The Normal Pass

This section describes the six-step sequence of questions and responses that constitutes the normal program flow of LABSRC3C. A direct progression through the six-step program flow is called a normal pass. However, additional commands may also be entered during the sequence with the exception of the search step Q04. Entering a command modifies the normal program flow, and turns control of the program over to the user. This chapter will be limited to a description of the normal program flow. The commands will be discussed separately in Chapter 3.

After the user requests LABSRC3C to be loaded, the main sequence of the program begins. Six questions are displayed in separate sequences for user response, as follows:

Q01 DO YOU WANT WORD ASSOCIATION?

Q02 PLEASE SPECIFY ASSOCIATION FILE.

Q03 DO YOU WANT SCORING?

Q04 ENTER BOOLEAN EXPRESSION.

Q05 DO YOU WANT RESULTS DISPLAYED?

Q06 SPECIFY RESTART OR EXIT.

The answers to these questions specify the kind of search that will be carried out. After typing a reply to each question, hit SEND BLOCK. This transmits the response to the computer, which will process it and then go on to the next question.

2.3 The Six Questions Individually Discussed

Q01 DO YOU WANT WORD ASSOCIATION?

(i.e. Do you want to search in associative mode?)

Valid Responses: YES
NO

A NO response instructs the system to search only for those documents whose index terms exactly match the specifications of the search request, to be submitted in response to Q04. This form of search is called direct match mode. If the response to Q01 is NO, the program will skip Q02 and go to Q03.

A YES response means that the system will search not only for documents whose index terms match those of your request, but also for documents indexed under terms, which, although not specified in the request, are highly associated (statistically) with the search request terms. This is called associative retrieval mode.

A YES response will therefore lead the system to retrieve all of the documents that would be produced from a NO response, and hopefully some others as well. For further details about term association, see Section 1.5 and Chapter 5.

Q02 PLEASE SPECIFY ASSOCIATION FILE

(i.e. Which association measure to you wish to use to expand your request?)

Valid Responses:
DOYLE
KUHNSG
KUHNSGN
KUHNSL
KUHNSS
KUHNSSN
KUHNSW
KUHNSY

An association file is a listing of each of the index terms in the Subject Authority List together with the four other index terms that are most closely associated (statistically) with it. The names of the different files correspond to different formulas for computing the closeness between pairs of index terms. The different methods or measures produce somewhat different results; therefore, the choice of association measure will affect both the quantity and the selection of the documents retrieved in response to your request. We do not yet fully understand how to choose from the repertory of association measures, although some suggestions are offered in Chapter 5. If you are uncertain or indifferent as to choice of file, we suggest KUHNSL.

Q03 DO YOU WANT SCORING?

(i.e. Do you want the program to display probable document relevance scores, when results are displayed?)

Valid Responses:
YES
NO

When scoring is specified, LABSRC3C computes a relevance score in the range (0,1) for each document, which reflects the closeness between the input request and the document retrieved. Thus, the relevance score is the degree to which the document indexing matches the request specification or its expansion. For an explanation of how the relevance score is determined, see Chapter 4.

If the response is NO, the system will not compute relevance scores, and the retrieval output will simply consist of an unranked set of document citations. If the search is being conducted in Direct Match Mode, i.e., the response to Q01 was NO, then it is of no advantage to reply YES to Q03, as all documents retrieved via direct match will have the same relevance number.

Q04 ENTER BOOLEAN EXPRESSION

Valid Responses: Any Boolean expression consisting of a string of terms from the Subject Authority List, connected by the operators AND, OR, or NOT.

Remarks on Q04

1. Each index term must be enclosed in single quotes; e.g.,

   'LANGUAGE' AND 'GRAMMAR' AND NOT 'SYNTAX'

2. Parenthetic expressions may be used; e.g.,

   ('AUTO. INDEXING' OR 'MANUAL INDEXING') AND
   (('SYSTEM' OR 'RETRIEVAL') AND 'EFFECTIVENESS')

3. Weights in the range (0,1) may be applied to elements of your Boolean expression. The weight is a 3-digit decimal number and must be followed by an asterisk; e.g.,

   'PERFORMANCE' OR .500*('PRECISION' OR 'RECALL')

   Warning: There is no diagnostic for a missing asterisk, but if the asterisk is not present, the system will ignore the weight in searching.

   For a discussion of the effects of weights on retrieval, see Chapter 4.

4. If the Boolean expression is longer than one line, LABSRC3C automatically concatenates the input. It will appear as if the last character of the line is dropped, but it will in fact be processed.

5. The response to Q04 must contain at least one Boolean operator. Although LABSRC3C is not intended for searching single-term requests, such requests may be formulated in the term-operator-term pattern by using the term as both of the operands; e.g.,

   'AUTOMATION' AND 'AUTOMATION'.

6. Commands may not be entered in response to Q04.

Q05 024 DOCUMENTS HAVE BEEN RETRIEVED

DO YOU WANT RESULTS DISPLAYED?

(i.e., Twenty-four documents have been retrieved by this request. Do you want to see accession numbers and relevance scores for these documents? Note: 024 is only an example. Any number of documents might have been retrieved.

Valid Responses: YES
NO

A YES response will generate a display of the accession numbers of the documents that satisfy the request. If the responses to both Q03 and Q05 were YES, then LABSRC3C will also display the probable relevance score for each retrieved document as well as the accession number. The relevance score will appear to the right of the accession number on the output display, as shown in Fig. 1.

If the response to Q05 is NO, the program will continue to

Q06 SPECIFY RESTART OR EXIT

Valid Responses: RESTART
EXIT

A RESTART response returns you to Q01.

EXIT causes LABSRC3C to be unloaded and returns control to the Terminal Monitor System. At this point another program may be requested (included LABSRC3C), or the session may be terminated. To sign off, the procedure is

hit CLEAR
hit TYPE
type LOGOUT
hit SEND BLOCK

AØØ28 .126	AØØ55 .Ø23	AØ131 .192	BØ5Ø2 .Ø36
BØ651 .Ø23	B1252 .192	XØØØ1 .996	XØØ13 .Ø24
XØØ14 .157	XØØ32 .996	XØØ65 .996	

FIG. 1

Example of a LABSRC3C output display in response to a search request in Associative Retrieval Mode, with scoring requested. Eleven documents have been retrieved, and are displayed in accession number order, together with their relevance numbers. Thus, for example, the first document that satisfies the request is A28, and its relevance number with respect to this request is .126.

3. COMMANDS

3.1 Introduction

The normal program flow described in Chapter 2 provides a basic automatic retrieval facility, but LABSRC3C also contains a command language which considerably increases the system's flexibility. These commands allow for altering the program flow, modifying requests, displaying selected portions of the list of retrieved documents, and displaying the MASTERI and MASTERA files and the association tables.

The components of a command in LABSRC3C are a verb and an object, although some commands use the verb alone. This verb-object combination is the basic pattern of the command. The program scans for this phrase in interpreting the input from the terminal. Other words may be added to the command if the user wants to enter a more complete English expression. For example:

```
GET 'B1218'
```

(the basic pattern) or

```
GET THE INDEX TERMS FOR DOCUMENT 'B1218'
```

will both cause the index terms for document B1218 to be retrieved, but use of the basic pattern minimizes the amount of typing required to communicate with the system.

Table 2 lists the basic patterns of the commands currently available in LABSRC3C in alphabetical order, briefly indicates their use, and serves as an index to this chapter.

3.2 GO TO--Commands for Changing the Normal Program Flow

The GO TO command in LABSRC3C causes the program to branch to the question indicated in the command. The basic pattern of this command is GO TO, followed by the number of the question to which you want the program to skip, e.g.,

```
GO TO Q04
```

If the GO TO command refers to a question number less than the number of the question currently displayed, the branch is referred to as a backward branch. Otherwise, it is a forward branch.

A forward branch enables you to skip over some questions. The system will then automatically substitute its own answers for the questions that you have skipped even if you have previously specified other answers.

These internally supplied answers are called default options. They are listed in Table 3.

Table 2: Index to the Commands

COMMAND	USE	EXAMPLE	DISCUSSION
DISPLAY			
(Association tables)	1. To see <u>association tables</u> for all terms in the request.	DISPLAY	3.3.1
	2. To see the association table for a particular term in the request.	DISPLAY 'EDUCATION'	3.3.1
(Document numbers with relevance scores, if scoring has been specified)	3. To see the retrieved set of <u>document nos. with relevance scores</u>.	DISPLAY DOCUMENTS	3.4.2
	4. To see a specified number of document nos. with relevance scores.	DISPLAY 4 DOCUMENTS	3.4.2
	5. To see document nos. with relevance scores for documents with scores greater than (GT), less than (LT) or equal to (EQ) a specified relevance score.	DISPLAY DOCUMENTS *GT*.500	3.4.2
	6. To see a specified number of documents GT, LT, or EQ a specified relevance score.	DISPLAY 3 DOCUMENTS *EQ*.460	3.4.2
EDIT	To display Boolean search request for editing.	EDIT	3.2.1
GET	1. To see the document <u>indexing</u> (MASTERI file record) for any document, whether or not it has been retrieved in the current search.	GET 'A0121'	3.4.1
	2. To see the MASTERI records for a specified number of <u>retrieved</u> documents.	GET 4 DOCUMENTS	3.4.2
GO TO	To <u>branch</u> to a different question.	GO TO Q02	3.1

Table 2: Index to the Commands (cont.)

COMMAND	USE	EXAMPLE	DISCUSSION
RETRIEVE	1. To see abstracts for all retrieved documents. 2. To see the abstract of a particular document whether or not it has been retrieved in the current search. 3. To see the abstracts of a specified number documents from the retrieved set. 4. To see abstracts for documents in the retrieved set with relevance scores greater than (GT), less than (LT) or equal to (EQ) a specified cut-off point. 5. To see abstracts for a specified number of the documents in the retrieved set GT, LT, or EQ a specified relevance score.	RETRIEVE DOCUMENTS RETRIEVE 'A0121' RETRIEVE 3 DOCUMENTS RETRIEVE DOCUMENTS *GT*.500 RETRIEVE 4 DOCUMENTS *LT*.200	3.4.3 3.4.2 3.4.2 3.4.2
SHOW	To see current parameters of the request, i.e., current answers to Q01-Q04.	SHOW	3.2.2
SORTA	To sort retrieved document numbers, MASTERI records or abstracts by relevance score in ascending ("lowest-first") order.	SORTA	3.3.2
SORTD	To sort retrieved document numbers, MASTERI records or abstracts by relevance score in descending ("highest-first") order.	SORTD	3.3.2

Table 3: Default Options

Question No.	Default Options
Q01 - DO YOU WANT WORD ASSOCIATION?	yes
Q02 - PLEASE SPECIFY ASSOCIATION FILE.	KUHNSG
Q03 - DO YOU WANT SCORING?	yes
Q04 - ENTER BOOLEAN EXPRESSION	previous Boolean expression
Q05 - DO YOU WANT RESULTS DISPLAYED?	no
Q06 - SPECIFY RESTART OR EXIT.	exit

Say, for example, that in response to

Q01 DO YOU WANT WORD ASSOCIATION

you type

GO TO Q04.

The system will internally establish the answers YES to Q01, KUHNSG to Q02, and YES to Q03, and the next message to be displayed will be

Q04 ENTER BOOLEAN EXPRESSION.

Q04 itself can not have a default option unless a Boolean expression has already been entered on a previous pass through the program; therefore, it makes no sense to execute a forward branch that skips over Q04 when you have just begun to run a search on LABSRC3C. Nevertheless, the default option for Q04 is one of the most powerful conveniences of the program because, once the request is submitted, you can change the association file and repeat the search without having to retype the request.

A backward branch is used for modifying a previous input and, therefore, will most likely be used in reply to Q06.

For instance, perhaps you have retrieved a set of documents by having the system search using the KUHNSG file and now would like to repeat the search on the KUHNSW file. After the system has displayed the results in the KUHNSG search, the next step in the program will be

Q06 SPECIFY RESTART OR EXIT.

To this you reply

GO TO Q02

and the system will respond with

Q02 SPECIFY ASSOCIATION FILE.

Your response will be

KUHNSW

and the system will proceed to

Q03 DO YOU WANT SCORING?

But the default option for this question is YES, and furthermore, you do not need to resubmit your Boolean expression. So instead of typing YES to Q03, which will cause the program to advance to Q04 and force you to retype your request, you type the command

GO TO Q05

and the system will now re-execute the search using your new association measure.

Note: There is an important difference between the inputs GO TO Q01 and RESTART. RESTART is a valid response to Q06 only, whereas GO TO Q01 may be sent in response to any question except Q04.

3.3 EDIT and SHOW--Commands for Modifying and Displaying the Boolean Request Expression and the Other Request Specifications

3.3.1 EDIT

This command enables you to change individual elements in a search request without having to retype the entire Boolean expression; thus, it is normally used as a reply to Q06. In response to the EDIT command, your previously typed Boolean expression will reappear on the screen, prefaced by the invitation EDIT AND THEN SEND BLOCK. At this point you may

- add or delete operands
- add or change Boolean operators
- replace any or all of your original operands with new ones
- add new weights
- change weights

These changes are made in the same way as error-correcting procedures for CRT inputs; i.e., by using the SPACE key to move the flashing cursor to positions on the terminal display screen corresponding to the characters you wish to delete, replace, or add. Once you have made the desired change in the Boolean expression, you complete the step in the usual way by hitting SEND BLOCK, and the system will now search for documents that satisfy your new request.

The EDIT command is especially useful for varying the elements of the Boolean expression one at a time to see what effect on

retrieval is produced by each of these changes. For example, suppose you wish to compare the set of document citations retrieved in response to the request

('LINGUISTIC' OR 'NATURAL LANGUAGE') AND 'ANALYSIS'

with the retrieval that will be obtained if the index term language is substituted for natural language. Use of the command GO TO Q04 will necessitate completely retyping the search request, just as if the previous one had never been submitted. Using EDIT instead can preserve the useful elements of the previous request. Instead of retyping the entire string, simply delete the word NATURAL from the original search expression to obtain the new one.

3.3.2 SHOW

This command provides a status report on your request, showing the current answers to questions Q01 to Q04. For the present, ignore the second "page" of this display. Its structure reflects distinctions which are not pertinent to present system implementation. SHOW is especially useful after you have been putting the system through a long series of request modifications and may not now be sure you remember all of the search specifications.

3.4 DISPLAY DOCUMENTS, SORTA, SORTD--Commands for Sorting, Selecting, and Displaying the Set of Retrieved Document Accession Numbers and Relevance Scores

3.4.1 DISPLAY DOCUMENTS

When DISPLAY DOCUMENTS is entered, document accession numbers and relevance values will be displayed on the CRT's. To call up a list of the documents that satisfy your input expression, you type

DISPLAY DOCUMENTS

This command, therefore, has the same effect as a YES response to Q05, but may be entered at any point in the normal program flow.

Its greatest advantage, however, is that it offers the opportunity to limit the length of the list of retrieved documents, an especially welcome capability when the number of documents which satisfy the input expression is inconveniently large. In such an event, you may request a restricted display by specifying the number of citations you wish to examine; e.g.,

DISPLAY 7 DOCUMENTS

Alternatively, you may also restrict the size of this list

by specifying a threshold relevance score or "cut-off" point:*

DISPLAY 10 DOCUMENTS *GT* .500

In this case the system response will depend on the number of documents that satisfy the restrictions specified in your command. If there are more than ten documents with relevance scores greater than .5, only ten will be displayed on the CRT's; but if there are less than ten documents satisfying the threshold requirement, then only that smaller number of documents will be displayed.

A third method for using DISPLAY to select portions of the file is to order the output using SORTD (Sec. 3.4.2) and then specify the number of documents desired. The use of these two commands will result in a display of the 5 documents with the highest relevance scores. Assume the input expression was ('AUTO. INDEXING' OR 'MANUAL INDEXING') AND 'INFO. RETRIEVAL'. The computer responds with the number of documents that satisfy the expression and Q05 - DO YOU WANT RESULTS DISPLAYED? You then type SORTD. The computer will respond with Q05 again. If you then type DISPLAY 5 DOCUMENTS, the program displays the five with the highest relevance scores.

3.4.2 SORTA and SORTD

The SORTA and SORTD commands are used to sort the documents by their relevance scores. SORTA command sorts them in ascending order; i.e., the documents with lowest relevance scores are listed first; SORTD sorts in descending order, producing a ranked output with the documents with the highest relevance scores listed first.

SORTA and SORTD are used only when scoring has been requested previously in response to Q03, and are usually entered in response to

Q05 DO YOU WANT RESULTS DISPLAYED?

SORTA and SORTD can also be used in combination with other commands such as DISPLAY to output selected portions of the set of retrieved documents as described in Sec. 3.4.1.

Similarly, instead of following SORTD with the DISPLAY command, you might prefer to type either

---

*The expression *GT* is the canonical abbreviation of the phrase "greater than." The command language also contains the expressions *EQ* and *LT*, "equal to" and "less than." These alternatives may be used in place of *GT* in any command that has *GT* as an element of its basic pattern.

GET 10 DOCUMENTS (See Sec. 3.5.2)

or

RETRIEVE 10 DOCUMENTS (See Sec. 3.5.3)

The use of SORTD in connection with GET and RETRIEVE is especially strategic, because GET and RETRIEVE both result in the output of a rather large amount of data about each document, and it can be tiresome to read through a long series of such representations.

3.5 GET, DISPLAY, and RETRIEVE--Commands for Displaying Data Files

3.5.1 GET

GET is used to display records from the MASTERI file, showing which index terms have been assigned to each document. There are two forms of the GET command. One of them calls for display of the indexing of a particular document; the other indicates the number of MASTERI records representing documents retrieved by the current request which are desired. To retrieve the indexing for a particular document, enter GET plus the document number; for example,

1. GET 'B1218'.

Note that the document number must consist of a letter followed by four digits, and be enclosed in single quotes.

Fig. 2 shows the display that will be provided in response to this command.

FIG. 2: MASTERI RECORD FOR DOCUMENT B1218

B121801	AUTO. INDEXING	CONNECTION	DESCRIPTOR	DOCUMENT
B121802	EVALUATION	INTRODUCTORY	MATCH	MATCH
B121803	PROBABILITY	QUESTION-ANSWER	RECALL	RELEVANCE
B121804	SEMANTIC	STATISTICAL	THESAURUS	VOCABULARY
B121804	WORD ASSOCIATION			

Any MASTERI record at all may be called up by this form of the GET command, regardless of whether it is a record for a document that satisfies a search request. In fact, you need not enter a Boolean expression at all in order to call for a display of a record from the MASTERI file.

The second form of the GET command does not require the specification of document numbers, but simply indicates the number of MASTERI records to be displayed. However, it selects documents only from the retrieved set. To see the documents in the retrieved set enter

2. GET DOCUMENTS

If, on the other hand, you don't want to see all the documents in the retrieved set, you can limit the number displayed. For example,

3. GET 5 DOCUMENTS

If the documents have been ordered by SORTD, this command will retrieve the five documents with the highest relevance score; otherwise, it retrieves them in accession number order.

3.5.2 DISPLAY

The use of DISPLAY to display the set of retrieved document accession numbers and relevance scores is discussed in Section 3.4.2. This section describes its use to display association tables.*

Assume that you have entered the request

('AUTO. INDEXING' AND 'MANUAL INDEXING') OR 'RETRIEVAL'

and you wish to know which terms are most highly associated with AUTO. INDEXING. To find out, you type:

1. DISPLAY 'AUTO. INDEXING'

The system will now respond by displaying the association table for AUTO. INDEXING, as shown in Fig. 3.

FIG. 3: TYPICAL ASSOCIATION TABLE

AUTO. INDEXING	9999
CRITICAL	.9300
BATCH PROCESSING	.8800
SCOPE NOTE	.4300
RELATIVE	.3586

*Association tables are defined in Sec. 1.6

The table reveals that the term most highly associated with AUTO. INDEXING is 'CRITICAL,' and that the association value of the pair of terms 'AUTO. INDEXING' and 'CRITICAL' is .9300.

If you wish to see the association tables for each term in your request, it is not necessary to specify a DISPLAY command for each individual term in your Boolean expression. Instead, you simply type:

2. DISPLAY

and the system will respond by displaying the association tables for each term in your Boolean expression, one at a time. When you are ready to have the next table displayed, hit SEND BLOCK.

3.5.3 RETRIEVE

RETRIEVE is used to display records from the MASTERA file, containing the author, title, and abstracts of the documents. There are four forms of this command.

1. RETRIEVE DOCUMENTS

The system will display the MASTERA records for all of the documents retrieved by the current search.

2. RETRIEVE DOCUMENTS *GT* .600

Here the system will display the abstracts of only those documents whose relevance number is greater than .600 (or whatever figure you specify in your command).*

3. RETRIEVE 5 DOCUMENTS

In this case the system will display the abstracts for the first 5 (or whatever figure you specify) documents retrieved in response to your request.

Remember that a YES response to Q05 causes the system to display document numbers in accession number order; hence the first 5 documents in the retrieval output will not necessarily be the 5 most relevant items. It is possible, however, to obtain a ranked output by responding to Q05 with a SORTD command instead of YES (See Sec. 3.4.2). If this sort has been done, the RETRIEVE command will display documents in order of decreasing relevance score.

---

*LT = (less than) and EQ = (equal to), may be substituted for GT in any command in which GT occurs, but remember that these expressions will not be understood by the program unless they are enclosed in asterisks; e.g., *LT*, *EQ*.

4. RETRIEVE 'B1218'

This command will have the system display the abstract of any particular document that you specify. The document number must be enclosed in single quotes and consist of one alphabetic character followed by four digits. The display that would be produced in response to this command is shown in Fig. 4.

The abstract of any document at all in the corpus may be called up by this form of the RETRIEVE command. The document need not be one of the ones retrieved in response to your search request.

The abstracts used for the MASTERA file have been for the most part taken from standard abstracting journals, such as Computing Reviews or Documentation Abstracts, and the source for each of the abstracts is given on the last line of the MASTERA record. In the example shown in Fig. 4, the word DOC on the last line signifies that the abstract that was printed as part of the document itself was the source for this particular MASTERA record.

FIG. 4: MASTERA RECORD FOR DOCUMENT B1218

```
B1218A1JA  PROBABILISTIC PAIRS AND GROUPS OF WORDS IN A TEXT
B1218A2AU  MEETHAM, A.R.
B121801AB  THE REPORT IS A CONTINUATION OF "PRELIMINARY STUDIES FOR MACHINE
B121802AB  GENERATED INDEX VOCABULARIES", A.R. MEETHAM (1963) LANGUAGE AND
B121803AB  SPEECH, 6,22.  IT ASSUMES THAT DOCUMENTS EMPLOY WORDS FROM PARTICULAR
B121804AB  GROUPS CONNECTED WITH THEIR SUBJECT MATTER, AND DISCUSSES FOUR
B121805AB  METHODS, TWO OF THEM NEW, FOR FINDING THE GROUPS.  THEY ARE PICKED
B121806AB  OUT FROM A WORD LIST BY USING A WORD-WORD BINARY MATRIX TO REPRESENT
B121807AB  THE ASSOCIATIONS BETWEEN PAIRS OF WORDS.  IN AN EVALUATION, THE
B121808AB  METHOD WHICH CONSUMES LEAST COMPUTER TIME TURNS OUT ALSO TO BE THE
B121809AB  BEST.
B121810AS  DOC
```

4. SCORING

4.1 General

Typing YES response to Q03 DO YOU WANT SCORING? causes LABSRC3C to compute a relevance score in the range (0, 1) for each document, which is one possible measure of the closeness of that document's index term set to your input request. The relevance scores are computed from the association values of the terms assigned to the documents. The actual computational procedure depends on the kinds of Boolean operators in the search request. These procedures are discussed one at a time in the following sections.

4.2 Requests with AND Operator

For a search in associative retrieval mode in which all the request terms are connected by AND operators, LABSRC3C will retrieve only those documents which have in their indexing a term from the association table of each operand in the original request. The expanded request is the original request plus the terms from the association tables added in associative mode to the original terms (see Sec. 1.6). The relevance score for the document is computed by multiplying the association values of the terms assigned to the documents that correspond to the terms in the expanded request. If more than one term from the same association table is present in a document's indexing and there is, therefore, more than one value to choose from for an operand, the highest value will be selected.

Example: Suppose the request

'CLASSIFICATION' AND 'CURRICULUM'

were submitted, and word association using the KUHNSL file and scoring requested. Word association will expand the request terms as follows:

Index Term	Assoc. Value	Index Term	Assoc. Value
CLASSIFICATION	.9999	CURRICULUM	.9999
LATTICE	.3191	EDUCATION	.6093
CATEGORIES	.2812	PHILOSOPHY	.5169
CLUMP	.2703	INFO. SCIENCE	.4496
PREDICTION	.2179	INTERDISCIPLINAR(Y)	.4439

Among the documents retrieved in response to this request, we find

A0049, "Librarianship and the Science of Information," by J.C. Donahue, American Documentation 17 (July 1966), 120-3 with a relevance score of .996, and

A0121, "Information Science and Liberal Education" by B.F. Cheydleur, American Documentation, 16 (July 1965), 171-7 with a relevance score of .170.

A0049 is indexed under the following terms:

CATALOGING	CLASSIFICATION	CURRICULUM	EDUCATION
INFO. SCIENCE	LIBRARIAN	PHILOSOPHY	

It was retrieved because both 'CLASSIFICATION' and 'CURRICULUM' are among its index terms, and these are precisely the terms specified by the request. From the table of association values, we see that 'CLASSIFICATION' and 'CURRICULUM' each have association values of .9999 with themselves. Since the terms were connected by the AND operator, the relevance score for A0049 will be computed by multiplying the association values:

.9999 × .9999 = .9998, which, however, comes out as

.996 due to founding properties of LABSRC3C's multiplication algorithm. Note that A0049 is also indexed under 'EDUCATION' and 'INTERDISCIPLINARY' which are all among the terms of the expanded request, since they are highly associated with 'CURRICULUM'. When there is more than one term in the expanded operand that is also in the indexing of a document, as mentioned above, LABSRC3C uses only the value for the term with the highest association value in performing the relevance computations. Since the association value of 'CURRICULUM' with itself (.9999) is, of course, at least as high as its association with any other term, and 'CURRICULUM' is present in the document's indexing, .9999 is used.

Document A0121 is indexed under the following terms:

CATEGORIES	CENTERS	CITATION INDEX	CODING
COMPUTER	DATA	DOCUMENT	EDITING
EDUCATION	IDENTIFICATION	INDEXING	INFO. RETRIEVAL
INFO. SCIENCE	INTERDISCIPLINAR(Y)	INTERFACE	LANGUAGE
LOGIC	MAN-MACHINE	MATHEMATICS	PROCESSING
QUESTION-ANSWER SYSTEM	SCOPE NOTE	STORAGE	STRUCTURE

This document is not indexed under 'CLASSIFICATION', but it is indexed under 'CATEGORIES', which has an association of .2812 with 'CLASSIFICATION'. Similarly, we do not find the second term of the request, 'CURRICULUM', among the terms which index A0121, but we do find 'EDUCATION', 'INFO. SCIENCE' and 'INTERDISCIPLINARY', which have association values of .6093, .4496 and .4439 respectively with 'CURRICULUM'. Since the indexing for A0121 contains at least one term which is highly associated with each of the request terms, this document was retrieved, even though there is no direct match between its index terms and those of the request. Its relevance score will be computed by multiplying the association value of the most closely associated terms, 'CATEGORIES' and 'EDUCATION', which are

.2812 × .6093 = .1713, which comes out as .170, due to rounding procedures in the computer multiplication program.

4.3 Requests with OR Operator

If all request terms are connected by OR operators, any document that has at least one of the request terms in its indexing will be retrieved. In searching in associative retrieval mode, the relevance score for the retrieved document will be computed as follows:

1) Consider the terms in the expanded request which also appear in the indexing of the document in question.

2) Select from these the term which has the highest association value with any of the original terms.

3) Assign the association value of that term as the relevance score of the document.

Example: Suppose the request is

'CLASSIFICATION' OR 'CLASSIF. SCHEME'

using the KUHNSL association measure. The association table for 'CLASSIFICATION' is, of course, the same as shown in Section 4.2. The table for 'CLASSIF. SCHEME' is

CLASSIF. SCHEME	.9999
MANUAL INDEXING	.3012
STAT. METHOD	.2562
TAG	.2380
CATEGORIES	.2380

Once again document A0121 will be retrieved. In this particular example there is a term in the indexing associated with both operands (although that is not necessary to retrieve the document), and, as frequently happens when the two operands are semantically as closely related as these two, it is the same term - CATEGORIES.

CATEGORIES is associated with CLASSIFICATION with an association value of .2812 and with 'CLASSIF. SCHEME' with a value of .2380. Since the scoring algorithm selects the higher value, the relevance value of A0121 relative to this request expression will be .2812.

4.4 Requests with NOT Operator

If a request term is preceded by the NOT operator, then no document indexed under that term will be retrieved, no matter how many other terms it may be indexed under that correspond to the non-negated specifications of the request. Thus the only effect that the NOT operator has on scoring is that the presence of the

negated term in the indexing of a document sends that document's relevance score to zero, and the document is not retrieved.

Negated terms are not expanded, even when searching in associative retrieval mode. There must be a direct match between a document's index term and a negated request term in order for the document to be rejected by virtue of the presence of the NOT operator in the request. The presence of a term associated with a negated term does not disqualify the document. So, for example, if there is a request of the form 'A' AND NOT 'B', and there is a document which is not indexed under either A or B, but is indexed under a term which is associated with both A and B, that document will be retrieved and assigned as a relevance score equal to that term's association value with term A.

Normally a term or any one of its associated terms will be considered in the search. Negating a term asks the system to check each document to make sure that the negated term does not appear; thus even though the term is negated, it is still included in the search. Therefore, when the association tables for a negated term are displayed, the term itself will not be followed by the asterisk which is used to indicate that a term is not currently being considered. The list of terms associated with the negated term, on the other hand, are not used to expand or restrict the request and thus are followed by asterisks in the association tables. (Note: if a term occurs in the list of a non-negated term as well as the negated one, it will be used to expand the non-negated term.)

Example: Consider a search in associative retrieval mode, carried out for the Boolean expression

'RELEVANCE' AND 'RECALL' AND NOT 'PRECISION'

The association tables for these terms will be displayed on the CRT's as follows:

RELEVANCE	RECALL	PRECISION
RECALL	PRECISION	RECALL *
MEASURE	CRANFIELD	CRANFIELD *
RELEV. JUDGMENT	EVALUATION	PERFORMANCE *
RANK	FACETED CLASSIFICATION	RANK *

This request is particularly interesting because one of the desired request terms, 'RECALL', is highly associated with the undesired term, 'PRECISION', and vice versa.

In this case the NOT operator takes precedence over the association data in the search. This is not to say that the request term 'RECALL' is not expanded: it means that 'RECALL' itself is expanded but that any document retrieved which also has been assigned the term 'PRECISION' will be excluded from the search result. Thus a document indexed under 'RELEVANCE'

and 'PRECISION' will not be retrieved, even though 'PRECISION' is highly associated with the desired term 'RECALL'. On the other hand, a document indexed under 'RELEVANCE' and 'CRANFIELD', but not indexed under 'PRECISION', will be retrieved, since 'PRECISION' is not expanded. The fact that 'CRANFIELD' is highly associated with the undesired term 'PRECISION' is ignored, and the document is retrieved and given a relevance number in the manner described in Sec. 4.2.

4.5 Weighting

4.5.1 The Effect of Weights on Relevance Scores

A weight may be any three digit decimal number from .000 to .999, and it is used to de-emphasize the term or the operand immediately following it in the Boolean expression. The weights may be applied to

(1) single term operands; i.e.,

.500*'AUTO. INDEXING' OR 'MANUAL INDEXING'

or (2) operands containing more than one term; i.e.,

.500*('AUTO. INDEXING' AND 'MANUAL INDEXING')
OR 'ABSTRACTING'

or (3) terms within operands; i.e.,

(.500*'AUTO. INDEXING' OR 'MANUAL INDEXING')
AND 'ABSTRACTING'.

An unweighted request term is automatically assigned a weight corresponding to its association value with itself.* When a different weight is explicitly assigned to a term in the Boolean expression, the association value of the term (as well as the association values of the four terms most highly associated with that term) is multiplied by the assigned weight. This operation will result in a lower relevance score for the documents that are retrieved because of the presence of this term in their indexing.

Example: Consider the request

'AUTO. INDEXING' OR 'MANUAL INDEXING'.

---

*Most of the files interpret the association value of a term with itself to be .999. This, however, is not true of KUHNSG, KUHNSS, and KUHNSW files. For an explanation of this phenomenon, see Chapter 5.

Assume that both of these terms have self-association values of .999. Thus a document indexed under 'MANUAL INDEXING' will be scored just as highly in the retrieval output as a document indexed under 'AUTO. INDEXING'. To de-emphasize the importance of 'MANUAL INDEXING', assign a weight of, for example, .500 to this term. The association value of 'MANUAL INDEXING' will now be recomputed as

$$.500 \times .999 = .499,$$

and the documents indexed under 'MANUAL INDEXING' will, therefore, have a lower relevance score and, in this case, be given a lower ranking in the retrieval output than the documents indexed under AUTO. INDEXING.

Note: Weights can only effect the ranking of the retrieved documents. They do not reduce the relevance score of a positively relevant document to zero. They do not change the selection of documents retrieved unless a cut-off point above .000 is also specified or unless, as sometimes happens, the relevance score is so low that the system rounds it off to .000.

4.5.2 The Effects of Term Weights on the Ranking of the Retrieved Set

Not all assignments of weights will, in fact, alter the ordering of the retrieved output although it will lower the relevance score of any document retrieved by the weighted term(s). If every document in the set, however, must contain the term or terms to be retrieved, then all document relevance scores will be proportionally affected by the weight. Specifically, one must be careful when using expressions which contain AND operators to see that the weighted term is not one which must appear on every retrieved document.

For example: In the following request, the weight may alter the ordering of the output:

'AUTO. INDEXING' AND (.500*'EVALUATION' OR 'EFFECTIVENESS')

since documents indexed under 'EVALUATION' and 'AUTO. INDEXING' will be affected by the weight, whereas those retrieved by 'AUTO. INDEXING' and 'EFFECTIVENESS' will rank above those retrieved by 'EVALUATION'.

Some examples of requests where the weights as shown will not alter the ordering of the output are:

1) .500*'AUTO ABSTRACTING' AND 'AUTO. INDEXING'

2) 'AUTO. INDEXING' and .500*('EVALUATION' OR 'EFFECTIVENESS')

Weights may be altered without retyping the whole request by using the EDIT command as described in Sec. 3.3.1.

5. ASSOCIATION FILES

5.1 Preliminaries

Chapter 5 is intended as a guide to the association measures available in LABSRC3C. The KUHNS and DOYLE measures are carefully derived and justified in their original presentations, and the reader is referred to these papers for a full understanding of their properties.* In this presentation, we will merely summarize their computational procedures and add a few interpretive comments.

5.2 Notation

Consider any pair of terms, A and B, from the Subject Authority List, used to index a corpus of N documents. Each document in the corpus can then be placed in one of four mutually exclusive and jointly exhaustive subsets of N.

A B: The subset of documents indexed under both A and B.
A $\overline{B}$: " " " " " " A but not under B.
$\overline{A}$ B: " " " " " " B " " " A.
$\overline{A}$ $\overline{B}$: " " " " " " neither A nor B.

Let: x = number of documents subset A B,
u = " " " " A $\overline{B}$,
v = " " " " $\overline{A}$ B,
y = " " " " $\overline{A}$ $\overline{B}$.

Furthermore, let n_1 = total number of documents indexed under A,
and n_2 = " " " " " " B.

The relationship between these expressions is summarized in Table 4.

---

*J.L. Kuhns, "The continuum of Coefficients of Association," in Statistical Association Methods for Mechanized Documentation, Symposium Proceedings, Washington, D.C., 1964, pp. 33-39. and

Lauren B. Doyle, "Indexing and Abstracting by Association," American Documentation 13: 378-390 (1962)

Table 4
Relationship Between Expressions A and B

	B	not-B	
A	x	$u = n_1 - x$	n_1
not-A	$v = n_2 - x$	$y = N - n_1 - n_2 + x$	$N - n_1$
	n_2	$N - n_2$	N

Following the customary notation of probability theory, we shall designate the probability of A as P(A), and the probability of A, given B, as $P(A|B)$.

5.3 Excess Over Independence Value

If terms A and B occur independently in the collection (i.e., if they are both randomly distributed throughout the file), one may calculate, based on their individual frequencies of occurrence, the number of documents in which one might expect the two terms to co-occur. This expected number of co-occurrences may be called the "independence value" since it is based upon the assumption that the two terms occur independently in the file.

The terms A and B are said to be statistically independent if $P(A|B) = P(A)$. In less formal language this means that if B is among the terms used to index a document, term A is neither more nor less likely to be among the index terms for that document than it would be if B were not present.

Using the notation of Table 4, P(A) is calculated by the ratio

$$\frac{n_1}{N}$$

and $P(A|B)$ is calculated as

$$\frac{x}{n_2} \quad .$$

So if A and B are statistically independent, the following equality holds:

$$\frac{x}{n_2} = \frac{n_1}{N} \quad .$$

every document in the collection. In such a case, $\delta(A, A) = 0$, which is just a numerical expression of the notion that a term applied to every document in the collection gives you no power to identify and retrieve suitable documents that you would not have by making a random selection from the collection.

5.4 Association Measures

The quantity $\delta(A, B)$ may range from large positive values to large negative values. To provide a common basis for comparison and calculation involving different pairs of index terms, some sort of normalizing factor is needed to bring the value of $\delta(A, B)$ into the range from -1 to +1, regardless of the values of x_1, n_1, and n_2.

There are, however, many normalizing factors that will bring about this result, and LABSRC3C offers a choice of such factors. When $\delta(A, B)$ is divided by one of these normalizing factors, the result is referred to as the coefficient of association. In other words:

$$\text{coefficient of association} = \frac{\text{excess over independence value}}{\text{normalizing factor}}$$

The numerical value of the coefficient of association for any pair of terms, A and B, is referred to as the association value of A with B. (See Sec. 1.6 for a more detailed definition.) The ratio is also referred to as the association measure, and the set of association values obtained by applying a specific association measure to all pairs of terms in the Subject Authority List is stored in an association file.

The coefficients of association used in LABSRC3C differ from each other only in the fact that they use different normalizing factors, (except for DOYLE, for which see Sec. 5.5.6). A summary of these measures is given in Table 5.

5.5 Notes on Individual Measures

We do not yet know enough about the behavior of the different measures to be able to make specific suggestions for a rational choice of a "best" measure for each kind of retrieval problem. Therefore, the descriptions that follow do not account for all properties of the measures, but rather represent explanations of commonly encountered problems in the interpretation of the association tables.

Notation: The expression S(A, B) is shorthand for "the association value between terms A and B as computed with the KUHNSS measure." Expressions like G(A, B), GN(A, B), etc. have similar meanings.

Multiplying both sides of this expression by n_2, we get

$$x = \frac{n_1 n_2}{N}$$

Now, subtracting $\frac{n_1 n_2}{N}$ from both sides of this equation, we obtain

$$x - \frac{n_1 n_2}{N} = 0 \qquad (1)$$

This equation will hold, if and only if A and B are independent. But the happy fact is that many pairs of index terms are not independent; for example, if a document is indexed under computers, it is more likely to also be indexed under automation than it would be if it were not indexed under computers. Thus, if we substitute the values of x, n_1 and n_2 representing the frequency of occurrence and co-occurrence of computers and automation in Eqn. (1), the quantity we obtain will not be zero, but rather some number greater than zero. This quantity is known as the excess over independence value, and is written in the Kuhns paper as

$$\delta(A, B).*$$

With a computer, it is easy to calculate the excess over independence value for every pair of terms in the Subject Authority List. In this way, we can determine, for any given term, the other terms which yield the highest values of $\delta(A, B)$. These terms are said to be the ones which are most statistically close to a given term.

Using this formula, the system also calculates a value for each term's association with itself, i.e., $\delta(A,A)$. In this case $x = n_1 = n_2$, and the left-hand side of Eqn. (1) will be modified as follows:

$$x - \frac{n_1 n_2}{N} = x - \frac{x^2}{N}$$

$$= x(1 - \frac{x}{N}) \qquad (2)$$

Eqn. (2) will yield higher values for terms which are used to index only a few documents than for terms which are used to index a great many. This means that, under this interpretation, not every term is equally close to itself. This notion is perhaps made clearer if we consider the extreme case where a term is used to index

*Kuhns also shows that this measure is symmetrical, i.e., for any pair of terms, A and B, $\delta(A, B) = \delta(B, A)$.

Table 5: Reference Chart of Kuhns Measures in LABSRC3C

In each case the expression in the right-hand column is divided into $\delta(A, B)$ to obtain the association value.

Name of Coefficient	Description	Normalizing Formula
KUHNSG	Angle between vectors	$\sqrt{n_1 n_2}$
KUHNSGN	Normalized* KUHNSG	$(1 - \frac{n_1}{N}\ \sqrt{n_1 n_2}$
KUHNSL	Linear correlation	$\sqrt{n_1 n_2 (1 - \frac{n_1}{N})\ (1 - \frac{n_2}{N}}$
KUHNSS	Area of separation	$N/2$
KUHNSSN	Normalized* KUHNSS	$n_1(1 - \frac{n_1}{N})$
KUHNSW	Conditional Probability	$\min(n_1, n_2)$
KUHNSY	Coefficient of colligation	$\frac{(\sqrt{xy} = \sqrt{uv})^2}{N}$

*Normalized in this context means weighted so that the association value of a term with itself under the measure is 1.0. In unmodified KUHNSG and KUHNSS measures the term's correlation with itself is less than 1.0.

In the discussion of the association tables, the letter A in such expressions as $\delta(A, B)$, $S(A, B)$, etc. will always stand for the header term.

5.5.1 KUHNSY

In this measure the association value of any term with itself is always 1,* and association values of highly associated terms tend to be quite high, very often greater than .8000. This is in particularly dramatic contrast to the KUHNSL measure, in which the coefficient of self-association is also 1, but where the coefficients of highly associated terms will be only about half the magnitude of the KUHNSY coefficients for the same pairs.

Theoretical digression: Strictly speaking, it is mathematically possible for $Y(A, B)$ to equal $L(A, B)$, rather than the usual case of $Y(A, B)$ being greater than $L(A, B)$. But in order for this to happen, both of the following conditions must be satisfied:

1. $u = v$
2. $x + y = 2\sqrt{xy}$

But if x and y are both integers, condition (2) can hold only if $x = y$, a condition which means that the number of documents indexed under both terms A and B equals the number of documents indexed under neither A nor B, a condition which is highly improbable in any "real" system.

Sometimes the results from the KUHNSY measures seem difficult to interpret, particularly when you ask the system to DISPLAY the association table for one of your request terms, and you get something that looks like this:

```
UPDATING        .9999
     WORD            .9999
       WEIGHT INDEXING      .9999
            WEIGHT              .9999
                 USER               .9999
```

This is the KUHNSY table of the terms most highly associated with updating, but the information seems to be capricious and counter-intuitive. What on eath have terms like word, weight-indexing, or weight got to do with updating? It takes a stretch of the imagination to conceive some reasonable connection, but even if one could be plausibly entertained, it seems too much to imagine a rational context in which the

*In the LABSRC3C association files, this quantity is "rounded" to .9999 for machine-computing convenience.

association between these terms would not only be stronger than any other imaginable terms, but also strong to the extraordinarily high degree of .9999!

The clue to this mystery lies in observing that the list of associated terms is in reverse alphabetical order. If the DISPLAY command could produce the ten terms most highly associated with updating instead of just four, this is what you would see:

UPDATING .9999
WORD .9999
WEIGHT INDEXING .9999
WEIGHT .9999
USER .9999
SYSTEM .9999
SIGNIFICANCE .9999
SELECTIVE DISSEM. .9999
SEARCHING .9999
PROFILE .9999
PROCESSING .9999

The reason for this peculiar state of affairs is this:

If, for any pair of co-occurring terms A and B, A never occurs without B, then the association value of A and B in the KUHNSY file will always be equal to 1 (i.e., .9999), regardless of how many times B occurs without A.

Restated in the notation of Sec. 5.2, this becomes

If $x = n_1$, $Y(A, B) = 1$.

Proof:

$$Y(A, B) = \frac{x - \frac{n_1 n_2}{N}}{(\sqrt{xy} + \sqrt{uv})^2/N}$$

From Table 4: $x + u = n_1$

But if $x = n_1$, then $u = 0$, and we can write:

$$Y(A, B) = \frac{x - \frac{xn_2}{N}}{(\sqrt{xy})^2/N}$$

But $n_2 = x + v$.

Making this substitution and multiplying both numerator and denominator by N for simplification, we obtain:

$$\frac{Nx - x(x+v)}{xy}$$

But N = x+y+v+u, and in the case under discussion u = 0; so this leads to the substitution:

$$\frac{x(x + y + v) - x(x + v)}{xy}$$

$$= \frac{x(x + y + v - x - v)}{xy}$$

$$= \frac{xy}{xy} = 1 = Y(A, B). \quad Q.E.D.$$

5.5.2 KUHNSW

The properties of this measure are a bit more clearly seen if we rewrite the W formula as the difference of two ratios. To simplify the notation, we will abbreviate $\min(n_1,n_2)$ and $\max(n_1,n_2)$ as "min" and "max," respectively.

$$W(A, B) = \frac{x - \frac{n_1 n_2}{N}}{\min}$$

$$= \frac{x - \frac{(\min)(\max)}{N}}{\min}$$

$$(3) \qquad = \frac{x}{\min} - \frac{\max}{N}$$

The B terms which will be most strongly associated with the given term, A, will be those which give highest value to this difference; in other words, the program will seek agreements which make x/min as small as possible and max/N as large as possible.

The first thing to note about the behavior of this measure is that max/N will ordinarily be small. Indeed, in a collection of any size, it will be negligible. Its usual role will be to rank terms with the same value of x/min, penalizing the more heavily posted terms. In a small experimental file, there may be a few cases where max/N is large enough to reduce a term's value below that of some terms having a lower value of x/min, but in general the rank of a term is determined by the value of x/min.

The highest possible value of x/min is 1, which is obtained whenever x=min, and, in particular, when computing W(A, A), A's association value with itself. In this case, as explained in Sec. 5.3.1, $x = \min = \max = n_1 = n_2$ and

$$(4) \qquad W(A, B) = 1 - \frac{\max}{N} = 1 - \frac{n_1}{N} = W(A, A)$$

If we rewrite W(A, A) as $1 - (1/N)\ n_1$, we can observe a straightforward relationship between a term's self-association value and its frequency of occurrence in the indexing of the collection. If there are 400 documents in the collection, then $1/N = .0025$. Thus, an index term that has been applied to one document will have a self-association value of .9975; a term that has been applied to two documents will have a value of .9950; and so forth.

It follows from this that the B terms which will be most highly associated with any given term A will be those whose frequencies are such that W(A, B) = W(A, A). This equivalence will be satisfied by any term B which <u>always</u> co-occurs with A (i.e. $x = n_2$). If A also always co-occurs with B, (i.e. $x = n_2 = n_1$), then Eqn. (3) reduces to W(A, A) as shown in Eqn. (4). The same result still obtains, however, whenever $x = n_2 = \min$ even if $n_1 \neq n_2$. In this case

$$(5) \qquad W(A, B) = \frac{x}{n_2} - \frac{n_1}{N} = 1 - \frac{n_1}{N} = W(A, A)$$

Thus, not only is $x = n_2$ a sufficient condition for a B term to be most highly associated with an A term, but, furthermore, the association value will be the same as the A term's self-association value. This is the explanation for outputs like this:

<u>Example</u>:

<u>Table</u>		<u>Frequency</u>
DOCUMENT	.6750	130
UPDATING	.6750	1
SUMMARY	.6750	2
SEE REFERENCE	.6750	1
SEE ALSO	.6750	2

This table is simply the top piece of a long list of lightly posted terms, which the program happens to display in reverse alphabetical order, all having the same maximum association value .6750 with 'DOCUMENT'. This maximum association value is low because document is such a heavily posted term.

The behavior of the measure is not so very different when $x = \min = n_1$; that is, when n_2 is the max. We have seen that when x = min, the maximum association values occur when the $\min = n_2$, but if $n_1 = 1$, or, at any rate, a small number, there may be no co-occurring terms with $x = \min$ such that $n_2 < n_1$. In this case the formula is:

(6) $$W(A, B) = 1 - \frac{max}{N} = 1 - \frac{n_2}{N}$$

But N is a constant; hence the smaller the value of n_2, the greater the value of W(A, B). Thus, the least frequently occurring terms will have the highest ranking. Furthermore, the interpretation of the association value is especially direct: n_2/N is simply the proportion of the collection to which term B has been assigned. Hence W(A, B) represents the proportion to which it has <u>not</u> been assigned. This is a thoroughly reasonable interpretation when we recall that KUHNSW is a conditional probability measure reflecting the probability of term B given that term A has occurred. It logically gives preference to terms which have the smallest number of occurrences without term A.

<u>Example</u>:

<u>Table</u>		<u>Frequency</u>
CRITICAL	.9975	1
EXTRACT	.9850	6
FALSE DROP	.9725	11
SURVEY	.9700	12
INFLUENCE	.9675	13

In this example, N = 400; hence $(1 - \frac{1}{N} = .9975$. Thus the association value for 'EXTRACT' was arrived at simply by multiplying .9975 by 6. Similarly, the other values are obtained by multiplying .9975 by n_2.

Now let us consider the case where x/min is less than 1. The most strongly associated terms will still be those for which (x - min) gives the smallest values. For a given value of x, call it x_0, such that x_0 > min, the highest value of W(A, B) will still be obtained when $n_1 = n_2$, because higher values of Eqn. (3) are obtained when max/N, the second element, is as small as possible. Since N is a constant, the smallest value of max/N occurs when the max = the min.

The most curious property of this measure, however, is that when $n_1 \neq n_2$, the measure gives higher rankings to terms whose frequency is such that n_2 will be the min. In other words, the measure favors terms which have fewer occurrences than the header term.

To see why this should be so, we need to refer again to Equ. (3). Since x/min is most likely to be near 1 when the min is small and max/N varies as the max, we see that the formula is most likely to yield high values when $n_1 + n_2$ is as small as possible. Since, for a given A, n_1 is a constant, the lowest

possible values of $n_1 + n_2$ occurs when n_2 is less than n_1. In other words, for a given value of x/min, the formula yields highest values when n_1 is the max. The note that follows will provide a more detailed explanation for the technically-minded reader.

Note: Consider two terms B_s and B_b which co-occur with A x_s and x_b times, respectively, and let the frequencies of B_s and B_b be such that $n_s < n_1 < n_b$. Substituting in Eqn. (3) for each of these two terms, we get

$$(7a) \qquad W(A_1B_s) = \frac{x_s}{n_s} - \frac{n_1}{N}$$

$$(7b) \qquad W(A_1B_b) = \frac{x_b}{n_1} - \frac{n_b}{N}$$

But N is a constant; hence $n_1/N < n_b/N$. Therefore, B_s will have a higher association value than B_b if $x_s/n_s \geq x_b/n_1$, but this inequality will always hold whenever $x_s \geq x_b$, regardless of the values of the other elements. This is simply an algebraic expression of the definition of KUHNSW as a conditional probability measure, favoring the terms which are less frequently posted than the header term.

To summarize:

1. Using KUHNSW, highest association values will be assigned to terms whose frequences are such that $x/n_2 = 1$. The most common case of this condition is when $n_2 = 1$; hence KUHNSW prefers terms with unique occurrences.

2. Next in ranking will be terms whose frequencies are such that $x < n_2 < n_1$, which provide the highest values of x/n_2, accounting for the most highly associated terms. Since higher values for x_1/n_2 are more likely when n_2 is small, KUHNSW will favor lightly posted terms.

Example:		Frequency	x	x/n_2
NETWORK	.9600	16	16	1
GOVERNMENT	.9600	1	1	1
VENN DIAGRAM	.6267	3	2	.67
PLANNING	.6267	3	2	.67
NATIONAL	.4600	6	3	.5

5.5.3 KUHNSS

The denominator of KUHNSS is a constant, and equal to half the number of documents in the collection. Thus this denominator is an order of magnitude larger than those of the other Kuhns measures, and the association values in the KUHNSS tables are rather small. For example, a pair of highly associated terms, like precision and recall, rate an association value of .1060 in the KUHNSS tables as compared to association values of .7571 and .6812 in the W and G tables, respectively.

Because the denominator of measure S is a constant, it serves as a normalizing factor only in that it prevents the magnitude of the coefficients from exceeding unity. It does not adjust the coefficients in such a way as to give all pairs of index terms an opportunity of having a high coefficient of association. Since the denominator of measure S is a constant, the relative size of two different coefficients will be determined entirely by the numerator. The numerator is dominated by the absolute number of co-occurrences of two terms. Two heavily posted terms have a much greater probability of co-occurring frequently than do two lightly posted terms. This means that the index terms that measure S finds highly associated with a heavily posted term will, themselves, be heavily posted. Thus, when a user's request involving a frequently assigned term is expanded according to measure S, several other heavily used terms are considered. This results in the retrieval of many more documents than if W or G were used.

An idiosyncracy of the S measure is that for some requests involving terms connected by the AND operator, associative retrieval may actually retrieve fewer documents than would be obtained by direct match. This paradox is explained in Sec. 5.5.5.

Although some users are dismayed by the S measure's propensity to retrieve somewhat voluminous and insensitively discriminated set of documents, if the SORTD command is invoked for a search in the S file, and only the documents with scoring greater than some suitable threshold are considered, the results are often unexpectedly gratifying. Although it seems reasonable that this should be so because the simple constant in the denominator preserves the relative magnitudes of the δ relationships between the various term pairs in the numerator, the full character of the S measure is not yet well understood. In general, however, it is the most suitable measure for searches where high recall is more important than high precision.

For the opposite case, where high precision is preferred to high recall, either of the measures discussed in the next section will, on the whole, be most likely to give satisfactory results.

5.5.4 KUHNSG and KUHNSL

These two measures are somewhat subtler than the other ones, and on the basis of present experience with, and understanding of them, they seem to be free of the obviously paradoxical or misleading properties of the other files.

The KUHNSL measure has the following virtues:

1. Self-association of a term is always .9999.

2. Term pairs which are highly associated will have "large" coefficients for their association values.

3. Perhaps with more consistency than the other measures, the most highly associated term-pairs in the L tables tend to show a recognizable correspondence to intuitive semantic plausibility. Therefore, associative searches with the L measure will in many cases be somewhat easier for less experienced users of LABSRC3C to interpret.

Most of these remarks also hold for KUHNSG, except that in this file self-association is not .9999, but is instead computed in the same way as with KUHNSW. KUHNSG tends to be "conservative," and for requests involving more than one AND operator it may only retrieve a few documents not already found on direct match. In such a case, a more extensive associative retrieval can be obtained by substituting the KUHNSGN measure. For further comments on KUHNSL, see Sec. 5.5.4.

5.5.5 The Normalized Measures, KUHNSGN and KUHNSSN

In the S and G measures, a term's self-association value is not equal to 1. Furthermore, the coefficient of association between two highly associated terms is less than, or equal to, a term's self-association value, and in fact may often be considerably less than the self-association value. For example, in the KUHNSG file the self-association of relevance is .8150, but the coefficient of relevance with its most highly associated term, recall, is only .2316.

As explained in Sec. 4.2, when a request which involves the AND operator is submitted to LABSRC3C, the relevance number of the documents is computed by multiplying the association values of the index terms. Since the association values are all less than one, the relevance number of a document indexed under three request terms in a request of the form 'A' and 'B' and 'C' will be a very small number indeed. However, LABSRC3C computes relevance numbers only to four places; thus if the product of the association values of the terms A, B, and C were some small value such as .00004, this will be interpreted by LABSRC3C as .0000, and the document will not be retrieved. This is particularly likely to happen when the request terms are weighted (see Sec. 4.5.1 for details of the computation of weighted terms).

In order to avoid failure to retrieve relevant documents under these conditions, LABSRC3C provides the two normalized files, KUHNSGN and KUHNSSN. In these files, a term's association with itself is set equal to one, and the values of associated terms are increased proportionally by dividing each value in the association table by the header term's self-association value. Here is an example from the KUHNSS tables:

OPTIMIZATION	.0149
HIERARCHY	.0085
STATISTICAL	.0082
EVALUATION	.0073
INFORMATION	.0069

When normalized for the KUHNSSN file, this table becomes

OPTIMIZATION	.0149/.0149 = 1 = .9999
HIERARCHY	.0085/.0149 = .5705
STATISTICAL	.0082/.0149 = .5503
EVALUATION	.0073/.0149 = .4899
INFORMATION	.0069/.0149 = .4631

As an example of the practical value of this device, consider a request of the form

'OPTIMIZATION' AND'SEARCH CRITERIA' AND 'RESPONSE TIME'

Now it seems reasonable that a document indexed under all three of these terms should be retrieved in response to this request. However, the self-association values of these terms in the S tables are:

OPTIMIZATION	.0149
SEARCH CRITERIA	.0099
RESPONSE TIME	.0099

and the scoring of a document indexed under all three of these terms will be .0149 × .0099 × .0099 = .0000015, which is too small a value for the document to be retrieved. But in the SN tables this scoring would be computed as $(.9999)^3 = .9996$, and the document would not only be retrieved, but also placed at the head of a ranked list of output citations, just as we would expect it to be.

The KUHNSGN file modifies the KUHNSG tables in a manner exactly analogous to the normalizing operation of the S tables discussed above. Unfortunately, the improved retrieval in this case is far less spectacular: on the average, $GN(A, B) \approx 1.2\ G(A, B)$.

5.5.6 DOYLE

This measure calculates the ratio of the logical product to the logical sum of the frequencies of pairs of index terms. In the notation of Sec. 5.2, this is expressed as

$$\text{DOYLE}(A, B) = \frac{x}{n_1 + n_2 - x}$$

Since DOYLE does not depend on N, it is especially sensitive to the ratio of n_1 and n_2, and is likely to assign larger coefficients to those term pairs where this ratio is close to 1.

Example: Say that $n_1 = 20$. In order for the coefficient to equal .2500, n_2 must be in the range (5, 80). The closer you get to the extreme of the range, the smaller the value $[\min(n_1, n_2)] - x$ must be in order to yield high values of this coefficient. But this is just to say that the closer the ratio of n_1/n_2 to 1, the smaller the ratio $x/\min(n_1, n_2)$ needs to be in order to obtain relatively high coefficients. But, by definition, this latter ratio cannot exceed 1, and the larger the value of $\min(n_1, n_2)$, the less the probability that $x = \min(n_1, n_2)$. See Fig. 4 on p. 24 for a display of this behavior.

When n_1 is very small, the probability that $x = n_1$ will be high. Under this condition, the Doyle equation reduces to x/n_2, and DOYLE will behave in the same way as Case 1 of the KUHNSW measure, as illustrated in Sec. 5.5.2. The value of the DOYLE coefficients will be smaller, but the selection of most highly associated terms will be the same.

FIG. 5:

DOYLE Coefficient of .2500 for $n_1 = 20$
As a Function of $\min(n_1, n_2)$

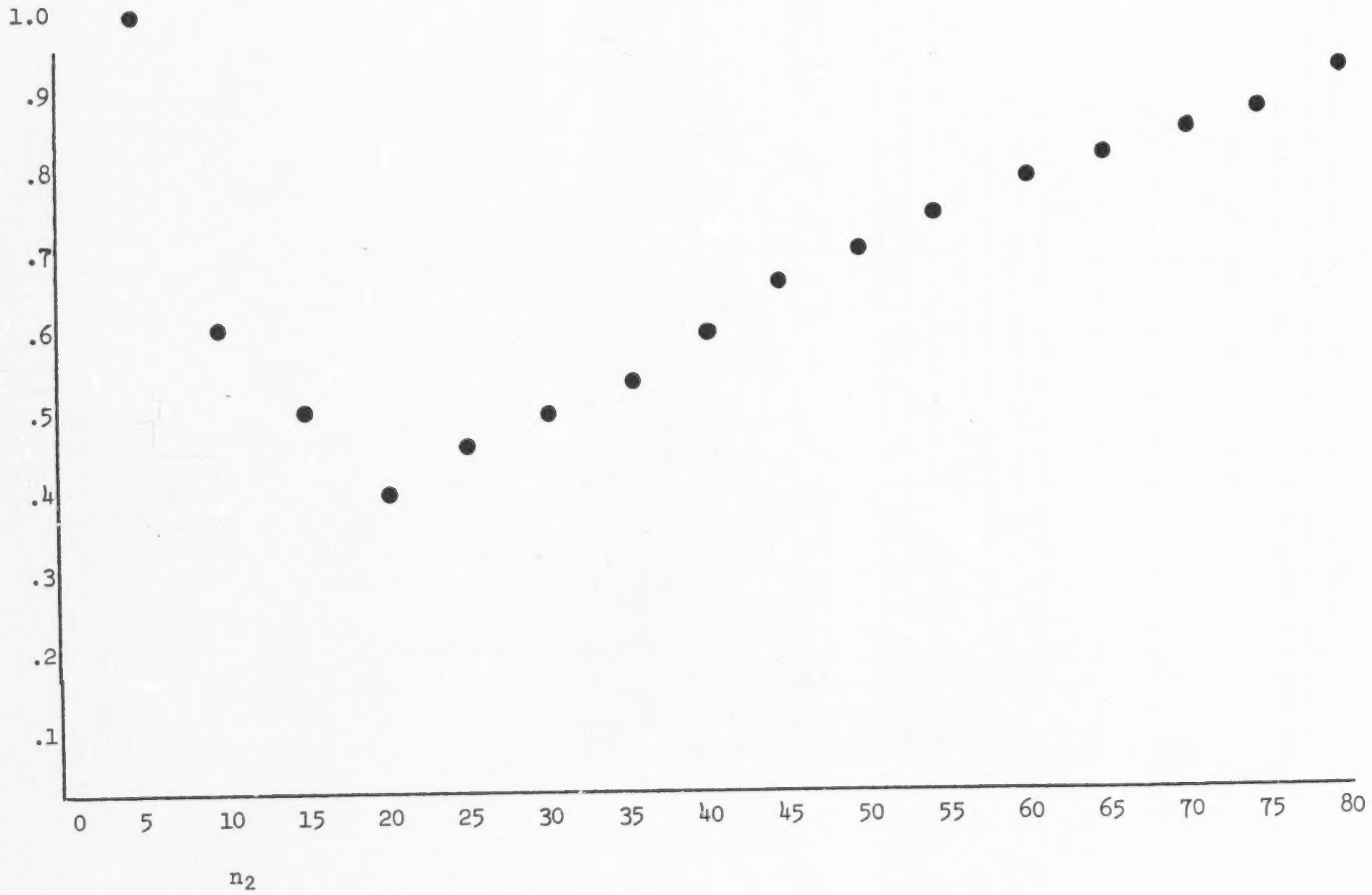

Table 6

Summarized Comparison of Doyle and Kuhns Measures

CONDITION	DOYLE	KUHNS
domain	does not depend on the value of N	depends on the value of N
range	(0, 1)	(-1, 1)
self-association of a term	always = 1	may be anywhere in the range (0, 1)
negative association	non negative	negative
exclusiveness	zero	negative
independence	positive	zero

INDEX OF KEY DEFINITIONS

	Section
accession number	1.4
association file	1.6
association measure,	
in general	1.6
in detail	ch. 5
association table	1.6
association value	1.6
associative retrieval mode	1.5
backward branch	3.2
basic patterns of commands	3.1
Boolean expressions, operators	
in general	1.7.1
in detail	ch. 4
coefficient of association	5.4
closeness, statistical	5.3
corpus	1.2
command	3.1
default option	3.2
direct match mode	1.5
expanded request	1.5
forward branch	3.2
independence value	5.3
MASTERA file	1.1
MASTERI file	1.1
normal program flow	2.2
operand	1.7
record	1.1
relevance number	4.1
relevance score	4.1
scoring	1.8
self-association	5.3.1
Subject Authority List	1.3
weights	3.3.1

APPENDIX 1

LABSRC3C SUBJECT AUTHORITY LIST

FILMED FROM BEST AVAILABLE COPY

SUBJECT AUTHORITY LIST

INFORMATION PROCESSING LABORATORY PROJECT
JANUARY 31, 1968
REVISED APRIL 20, 1969

ABBREVIATIONS

S = SEE
SA = SEE ALSO
SN = IN THE SENSE OF (I.E. SCOPE NOTE)
* = NO DOCUMENTS YET INDEXED WITH THIS TERM
+ = TERM NOT ALLOWED, RELATED TERM TO BE USED

*ABBREVIATION
ABSTRACT
ABSTRACTING
ACCESS
ACCESSION NUMBER
ACCURACY
ACQUISITION
ACTIVITY
ADDRESS
ADMINISTRATION
ALGEBRA
+ALGOL
S PROG. LANGUAGE
ALGORITHM
ALPHABETIC
ALPHABETIC ORDER
ALPHANUMERIC
*ALTERNATIVES
AMBIGUITY
ANALOGY
ANALYSIS
ANSWER
*ANTHOLOGY
SA BIBLIOGRAPHY
APPLICATION
+ARITHMETIC
S MATHEMATICS
ARRAY
+ARTICLE
S DOCUMENT
ARTIFICIAL INTEL
ARTIFICIAL LANG.
ASSIGNED

ASSOCIATION
SA WORD ASSOCIATION
ASSOCIATIVE
+ATTRIBUTE
S CHARACTERISTIC
AUTHOR
AUTHORITY LIST
SA THESAURUS
AUTO ABSTRACTING
AUTO. INDEXING
AUTOMATIC
AUTOMATION
SA MECHANIZATION

BATCH PROCESSING
BEHAVIORAL
BIBLIOGRAPHIC
BIBLIOGRAPHY
SA ANTHOLOGY
BINARY
BOOK
BOOLEAN
SA LOGICAL
BROWSING

CALL NUMBER
CANONICAL
SA NORMALIZED
CARD
CARD CATALOG
CATALOG

FILMED FROM BEST AVAILABLE COPY

CATALOGING
CATEGORIES
CENTERS
CENTRALIZED
CHANNEL
CHARACTERISTIC
CHEMICAL
CIRCULATION
CITATION
CITATION INDEX
*CLAIM
SA COPYRIGHT
SA PATENT
CLASSIF. SCHEME
CLASSIFICATION
CLERICAL
+CLUE WORD
S KEYWORD
CLUMP
CLUSTER
CO-OCCURRENCE
+COBOL
S PROG. LANGUAGE
CODE
SN MEDIA DESIGNATION
CODING
SN COMPUTER CODING
COEFFECIENT
COLLECTION
*COLLOQUIUM
SA CONFERENCE
SA MEETING
SA SYMPOSIUM
+COLON CLASSIF.
S FACETED CLASSIF.
COMBINATIONS
+COMIT
S PROG. LANGUAGE
COMMUNICATION
COMP LINGUISTICS
COMPARISON
COMPONENT
COMPUTER
CONCEPT
CONCORDANCE
CONDITIONAL PROB
CONFERENCE
SA COLLOQUIUM
SA MEETING
SA SYMPOSIUM
CONNECTION
+CONSECUTIVE
S ORDER
+CONSOLE
S REMOTE TERMINAL
CONTENT
CONTENT ANALYSIS
CONTEXT
CONTROL
CONTROLLED
CONVENTIONAL
CONVERSION
COORDINATE
COORDINATE INDEX
SA UNITERM SYSTEM
*COPYRIGHT
SA CLAIM
SA PATENT
+CORE
S STORAGE
CORRELATION
COST
COUNT
COUPLING
CRANFIELD
CRITERIA
CRITICAL
SN REVIEWING, NOT VITAL
CROSS REFERENCE
CURRENT AWARENES
CURRICULUM
+CUSTOMER
S USER
CYBERNETICS

DATA
DECENTRALIZATION
DECISION THEORY
DEDUCTIVE
DEGREE
DEPTH OF INDEXIN
DESCRIPTIVE
DESCRIPTOR
SA KEYWORD
SA TAG
SA TERM
DESIGN
SA PLANNING
DICTIONARY
+DIFFERENCE
S COMPARISON
+DIGITAL COMPUTER
S COMPUTER
DISCRIMINANT
+DISPLAY
S REMOTE TERMINAL
DISSEMINATION
*DISSERTATION
DOCUMENT
SA JOURNAL
DOCUMENTATION

FILMED FROM BEST AVAILABLE COPY

DUAL DICTIONARY

+ECONOMICS
S COST
EDITING
EDUCATION
EFFECTIVENESS
SA EFFICIENCY
EFFICIENCY
SA EFFECTIVENESS
+ELECTRONIC COMPUTER
S COMPUTER
+ELEMENT
S COMPONENT
+EMPIRICAL
S EXPERIMENT
+ENCODING
S CODING
+ENGLISH
S NATURAL LANGUAGE
ENTROPY
ENTRY
SN ACCESS POINT
+EQUIPMENT
S HARDWARE
ERROR
EVALUATION
SA TEST
SA UTILITY
SA VALUE
EXPERIMENT
EXTRACT

FACET
FACETED CLASSIF.
FACT RETRIEVAL
+FACTOR ANALYSIS
S STAT. METHOD
FALSE DROP
FEEDBACK
FILE
SA LIST
SA STRING
FILE ORGANIZATIO
FLOW OF INFO.
FORMAT
+FORTRAN
S PROG. LANGUAGE
FREQUENCY
SA WORD FREQUENCY
FUNCTION
SN OPERATIONAL, NOT MATHMATICAL

GENERAL
GENERATION
SN PRODUCTION
GENERIC
+GOAL
S OBJECTIVE
GOVERNMENT
GRAMMAR
GRAPH
SN MATHEMATICAL GRAPH
SA TABLE
GRAPHICS
SN GRAPHIC MATERIALS E.G. PHOTOS.
+GROUP
S CLUMP

HARDWARE
SN COMPUTERS, MICROFILM EQUIPMENT, ETC.
SA MECHANICAL
+HEADINGS
S SUBJECT HEADING
HIERARCHY
HISTORICAL
+HUMAN
S MANUAL
+HUMAN INDEXING
S MANUAL INDEXING

*IDENTICAL
IDENTIFICATION
ILLUSTRATION
*IMPLEMENTATION
INDEPENDENT
INDEX
+INDEX TERM
S TERM
INDEXING
SA SUBJECT INDEXING
INFERENCE
+INFO. LANGUAGE
S ARTIFICIAL LANG.
INFO. RETRIEVAL
INFO. SCIENCE
INFORMATION
+INFORMATION FLOW
S FLOW OF INFO.
INPUT
+INQUIRER
S USER
+INQUIRY
S QUESTION

FILMED FROM BEST AVAILABLE COPY

+INSTRUCTION
S EDUCATION
INTELLECTUAL
INTERACTION
INTERDISCIPLINAR
INTERFACE
INTERPRET
+INTERROGATE
S QUESTION
+INTERSECTION
S VENN DIAGRAM
INTRODUCTORY
INTUITIVE
INVENTORY
INVERTED
IRRELEVANT
+ITEM
S DOCUMENT
ITERATIVE
SA RECURSIVE

JOURNAL
SA DOCUMENT

KEYPUNCH
KEYWORD
SA DESCRIPTOR
SA TAG
SA TERM
+KNOWLEDGE
S INFORMATION
KWIC

LABORATORY
LANGUAGE
SA ARTIFICIAL LANG.
SA NATURAL LANGUAGE
LARGE
LATTICE
LAW
+LEVEL
S DEGREE
+LEXICAL
S ALPHABETIC
+LEXICON
S DICTIONARY
LIBRARIAN
LIBRARY
LINEAR
LINGUISTIC
LINK
LIST
SA FILE
SA STRING
LITERATURE
LOGIC
LOGICAL
SA BOOLEAN

+MACHINE
S HARDWARE
MACHINE-READABLE
+MAGNETIC TAPE
S STORAGE
MAN-MACHINE
MANUAL
MANUAL INDEXING
MATCH
MATHEMATICAL
MATHEMATICS
SA PROBABILITY
MATRIX
MEANING
MEASURE
MECHANICAL
SA HARDWARE
MECHANIZATION
SA AUTOMATION
MEDIUM
MEETING
SA COLLOQUIUM
SA CONFERENCE
SA SYMPOSIUM
+MEMORY
S STORAGE
MESSAGE
METHODOLOGY
+METRIC
S MEASURE
MICROFICHE
MICROFILM
MODEL
SA SIMULATION
MODIFICATION
MORPHOLOGY
MULTIPLE

NATIONAL
NATURAL
NATURAL LANGUAGE
NEEDS
NETWORK
SN ORGANIZATIONAL STRUCTURE
SA ORGANIZATION
NOISE

FILMED FROM BEST AVAILABLE COPY

+NOMENCLATURE
S NOTATION
NON-CONVENTIONAL
NON-DISCRIMINANT
NON-FILE
NON-RANDOM
NON-RELEVENT
*NORMALIZED
SA CANONICAL
NOTATION
SA TERMINOLOGY
NUMBER
NUMERIC

OBJECTIVE
SN GOAL, NOT AS OPPOSED TO SUBJECTIVE
OCCURRENCE
OFF-LINE
ON-LINE
OPERATION
+OPTICAL COINCIDENCE
S PEEK-A-BOO
OPTIMIZATION
ORDER
ORGANIZATION
SA NETWORK
OUTPUT

+PAIR
S WORD ASSOCIATION
+PAPER
S DOCUMENT
PARAMETER
SA VARIABLE
PARSE
PATENT
SA CLAIM
SA COPYRIGHT
PATTERN
PEEK-A-BOO
PERFORMANCE
+PERIODICAL
S JOURNAL
PERMUTED
PERTINENT
SA RELEVANT
PHILOSOPHY
SA POLICY
+PHOTO
S GRAPHICS
PHRASE
PLANNING
SA DESIGN

+PLOT
S GRAPH
+POLICY
SA PHILOSOPHY
+POPULATION
S COLLECTION
PRECISION
PREDICTION
*PRINCIPLE
+PRINT-OUT
S OUTPUT
PRINTING
+PRIVACY
S SECRECY
PROBABILITY
SA MATHEMATICS
PROCEDURE
PROCEEDINGS
PROCESSING
PROFILE
PROG. LANGUAGE
PROGRAM
SN COMPUTER PROGRAM
SA ROUTINE
SA SOFTWARE
SA SUBROUTINE
PROGRAMMED
+PROPERTY
S CHARACTERISTIC
PSYCHOLOGY
+PUBLICATION
S DOCUMENT
PUNCHED
+PUNCHED-CARD
S STORAGE
PUNCTUATION
+PURPOSE
S OBJECTIVE

QUALITATIVE
SA SUBJECTIVE
QUANTITATIVE
+QUERY
S QUESTION
QUESTION
SN BOTH NOUN AND VERB
QUESTION NEGOT.
QUESTION-ANSWER

RANDOM
RANDOM-ACCESS
RANK
RATE
READING

FILMED FROM BEST AVAILABLE COPY

REAL-TIME
RECALL
RECOGNITION
RECORD
+RECORDED INFO.
S RECORD
RECURSIVE
SA ITERATIVE
REDUNDANCY
REFERENCE
*REJECTION
RELATED
RELATIONSHIP
RELATIVE
RELEV. JUDGEMENT
RELEVANCE
RELEVANT
SA PERTINENT
+REMOTE TELETYPES
S REMOTE TERMINAL
REMOTE TERMINAL
SA VISUAL DIS. CON.
+REPORT
S DOCUMENT
+REQUEST
S QUESTION
RESEARCH
+RESPONSE
S ANSWER
RESPONSE TIME
RETRIEVAL
RETRIEVAL SYSTEM
REVIEW
SA SUMMARY
SA SURVEY
ROLE
ROUTINE
SN COMPUTER ROUTINE
SA PROGRAM
SA SOFTWARE
SA SUBRCUTINE
RULE

SAMPLE
SCANNING
SCIENTIFIC
SCOPE NOTE
SEARCH CRITERIA
SEARCH STRATEGY
SEARCHING
*SECRECY
SEE ALSO
SN AS USED IN CATALOGING
SEE-REFERENCE
SELECTION

SELECTIVE DISSEM
SEMANTIC
SA SYNTAX
SEQUENCE
+SEQUENTIAL
S LINEAR
+SERIAL
S JOURNAL
SERVICE
SET THEORY
SETS
SHELFLIST
SIGNIFICANCE
SIMULATION
SA MODEL
SIZE
SMALL
SMART SYSTEM
SOCIAL IMPLIC.
SOFTWARE
SA PROGRAM
SA ROUTINE
SA SUBROUTINE
SORTING
SOURCE
SPECIALIZED
SPECIFICITY
STANDARDIZATION
STAT ASSOCIATION
STAT. ANALYSIS
SA STAT. METHOD
STAT. METHOD
SA STAT. ANALYSIS
STATE-OF-THE-ART
STATISTICAL
+STOCHASTIC
S RANDOM
STORAGE
STRING
SA FILE
SA LIST
STRUCTURE
SUBJECT
SUBJECT HEADING
SUBJECT INDEXING
SUBJECT-CATALOG.
SUBJECTIVE
SA QUALITATIVE
SUBROUTINE
SA PROGRAM
SA ROUTINE
SA SOFTWARE
SUMMARY
SA REVIEW
SA SURVEY

FILMED FROM BEST AVAILABLE COPY

SURVEY
SA REVIEW
SA SUMMARY
SYMBOL
SYMBOLIC LOGIC
SYMPOSIUM
SA COLLOQUIUM
SA CONFERENCE
SA MEETING
SYNONYM
SYNTACTIC ANAL.
SYNTAX
SA SEMANTIC
SYSTEM

TABLE
SA GRAPH
TAG
SA DESCRIPTOR
SA KEYWORD
SA TERM
+TAPE
S STORAGE
+TEACHING
S EDUCATION
TECHNICAL
TECHNICAL REPORT
TECHNOLOGY
TELEGRAPHIC ABS.
TERM
SA DESCRIPTOR
SA KEYWORD
SA TAG
+TERMINAL
S REMOTE TERMINAL
TERMINOLOGY
SA NOTATION
TEST
SA EVALUATION
SA UTILITY
SA VALUE
TEXT
THEORY
THESAURUS
SA AUTHORITY LIST
TIME
TIME-SHARING
TITLE
+TOPIC
S SUBJECT
TRANSFORMATION
TRANSLATION
*TRANSLITERATION
TRANSMISSION
TREE
TREE STRUCTURE
TRUNCATION
*TYPE STYLE
TYPE-SETTING
*TYPOGRAPHICAL

+UNION
SN SET THEORY UNION
S VENN DIAGRAM
*UNION CATALOG
+UNITERM
S DESCRIPTOR
UNITERM SYSTEM
SA COORDINATE INDEX
UPDATING
USER
USER STUDY
UTILITY
SA EVALUATION
SA TEST
SA VALUE

VALIDATION
VALUE
SA EVAUATION
SA TEST
SA UTILITY
VARIABLE
SA PARAMETER
VECTOR
VENN DIAGRAM
*VISUAL DIS. CON.
SA REMOTE TERMINAL
VOCABULARY

WEIGHT
WEIGHT INDEXING
WORD
WORD ASSOCIATION
WORD FREQUENCY
+WORD PAIRS
S WORD ASSOCIATION

FILMED FROM BEST AVAILABLE COPY

APPENDIX 2

LIST OF INDEX TERM FREQUENCIES IN THE CORPUS

FILMED FROM BEST AVAILABLE COPY

INDEX TERM LIST ALPHABETICALLY SORTED. LIST REVISED APRIL 24, 1969

INDEX TERM	NO. OF REFS.
ABBREVIATION	00000
ABSTRACT	00027
ABSTRACTING	00013
ACCESS	00016
ACCESSION NUMBER	00004
ACCURACY	00005
ACQUISITION	00005
ACTIVITY	00001
ADDRESS	00010
ADMINISTRATION	00001
ALGEBRA	00014
ALGORITHM	00035
ALPHABETIC	00006
ALPHABETIC ORDER	00001
ALPHANUMERIC	00002
ALTERNATIVES	00000
AMBIGUITY	00006
ANALOGY	00002
ANALYSIS	00076
ANSWER	00010
ANTHOLOGY	00000
APPLICATION	00004
ARRAY	00004
ARTIFICIAL INTEL	00002
ARTIFICIAL LANG.	00006
ASSIGNED	00002
ASSOCIATION	00050
ASSOCIATIVE	00017
AUTHOR	00006
AUTHORITY LIST	00003
AUTO ABSTRACTING	00014
AUTO. INDEXING	00028
AUTOMATIC	00055
AUTOMATION	00010
BATCH PROCESSING	00001
BEHAVIORAL	00002
BIBLIOGRAPHIC	00023
BIBLIOGRAPHY	00015
BINARY	00008
BOOK	00009
BOOLEAN	00020
BROWSING	00004
CALL NUMBER	00001
CANONICAL	00005
CARD	00009
CARD CATALOG	00002
CATALOG	00009
CATALOGING	00007
CATEGORIES	00016
CENTERS	00007
CENTRALIZED	00003
CHANNEL	00003

FILMED FROM BEST AVAILABLE COPY

CHARACTERISTIC	00011
CHEMICAL	00003
CIRCULATION	00004
CITATION	00013
CITATION INDEX	00013
CLAIM	00000
CLASSIF. SCHEME	00021
CLASSIFICATION	00000
CLERICAL	00004
CLUMP	00022
CLUSTER	00020
CO-OCCURRENCE	00026
CODE	00010
CODING	00020
COEFFICIENT	00024
COLLECTION	00015
COLLOQUIUM	00000
COMBINATIONS	00004
COMMUNICATION	00036
COMP LINGUISTICS	00006
COMPARISON	00027
COMPONENT	00003
COMPUTER	00112
CONCEPT	00020
CONCORDANCE	00003
CONDITIONAL PROB	00007
CONFERENCE	00006
CONNECTION	00008
CONTENT	00024
CONTENT ANALYSIS	00015
CONTEXT	00010
CONTROL	00003
CONTROLLED	00003
CONVENTIONAL	00003
CONVERSION	00003
COORDINATE	00009
COORDINATE INDEX	00032
COPYRIGHT	00000
CORRELATION	00021
COST	00022
COUNT	00003
COUPLING	00008
CRANFIELD	00010
CRITERIA	00012
CRITICAL	00001
CROSS REFERENCE	00009
CURRENT AWARENESS	00009
CURRICULUM	00011
CYBERNETICS	00002
DATA	00034
DECENTRALIZATION	00001
DECISION THEORY	00003
DEDUCTIVE	00004
DEGREE	00010
DEPTH OF INDEXIN	00011
DESCRIPTIVE	00007
DESCRIPTION	00030

FILMED FROM BEST AVAILABLE COPY

DESIGN	00022
DICTIONARY	00044
DISCRIMINANT	00005
DISSEMINATION	00012
DISSERTATION	00000
DOCUMENT	00130
DOCUMENTATION	00013
DUAL DICTIONARY	00001
EDITING	00009
EDUCATION	00019
EFFECTIVENESS	00015
EFFICIENCY	00029
ENTROPY	00004
ENTRY	00007
ERROR	00014
EVALUATION	00072
EXPERIMENT	00053
EXTRACT	00006
FACET	00007
FACETED CLASSIF.	00008
FACT RETRIEVAL	00005
FALSE DROP	00011
FEEDBACK	00014
FILE	00033
FILE ORGANIZATIO	00014
FLOW OF INFO.	00007
FORMAT	00006
FREQUENCY	00038
FUNCTION	00015
GENERAL	00009
GENERATION	00004
GENERIC	00013
GOVERNMENT	00001
GRAMMAR	00031
GRAPH	00024
GRAPHICS	00001
HARDWARE	00020
HIERARCHY	00039
HISTORICAL	00005
IDENTICAL	00000
IDENTIFICATION	00002
ILLUSTRATION	00010
IMPLEMENTATION	00000
INDEPENDENT	00001
INDEX	00054
INDEXING	00097
INFERENCE	00013
INFO. RETRIEVAL	00105
INFO. SCIENCE	00015
INFORMATION	00083
INPUT	00037
INTELLECTUAL	00004
INTERACTION	00005
INTERDISCIPLINAR	00007
INTERFACE	00003
INTERPRET	00008
INTRODUCTORY	00008

FILMED FROM BEST AVAILABLE COPY

INTUITIVE	00002
INVENTORY	00002
INVERTED	00002
IRRELEVANT	00006
ITERATIVE	00004
JOURNAL	00008
KEYPUNCH	00003
KEYWORD	00023
KWIC	00013
LABORATORY	00003
LANGUAGE	00059
LARGE	00002
LATTICE	00018
LAW	00006
LIBRARIAN	00014
LIBRARY	00036
LINEAR	00001
LINGUISTIC	00036
LINK	00021
LIST	00018
LITERATURE	00026
LOGIC	00026
LOGICAL	00010
MACHINE-READABLE	00006
MAN-MACHINE	00017
MANUAL	00009
MANUAL INDEXING	00002
MATCH	00030
MATHEMATICAL	00020
MATHEMATICS	00009
MATRIX	00054
MEANING	00020
MEASURE	00043
MECHANICAL	00012
MECHANIZATION	00017
MEDIUM	00004
MEETING	00005
MESSAGE	00003
METHODOLOGY	00019
MICROFICHE	00004
MICROFILM	00003
MODEL	00038
MODIFICATION	00001
MORPHOLOGY	00003
MULTIPLE	00003
NATIONAL	00006
NATURAL	00002
NATURAL LANGUAGE	00059
NEEDS	00015
NETWORK	00016
NOISE	00011
NON-CONVENTIONAL	00002
NON-DISCRIMINANT	00001
NON-FILE	00001
NON-RANDOM	00001
NON-RELEVANT	00003
NORMALIZED	00006

FILMED FROM BEST AVAILABLE COPY

NOTATION	00015
NUMBER	00002
NUMERIC	00006
OBJECTIVE	00004
OCCURRENCE	00003
OFF-LINE	00002
ON-LINE	00007
OPERATION	00005
OPTIMIZATION	00003
ORDER	00013
ORGANIZATION	00012
OUTPUT	00031
PARAMETER	00011
PARSE	00014
PATENT	00006
PATTERN	00006
PEEK-A-BOO	00003
PERFORMANCE	00028
PERMUTED	00008
PERTINENT	00006
PHILOSOPHY	00003
PHRASE	00007
PLANNING	00003
PRECISION	00028
PREDICTION	00006
PRINCIPLE	00006
PRINTING	00001
PROBABILITY	00039
PROCEDURE	00013
PROCEEDINGS	00008
PROCESSING	00042
PROFILE	00008
PROG. LANGUAGE	00013
PROGRAM	00041
PROGRAMMED	00004
PSYCHOLOGY	00003
PUNCHED	00005
PUNCTUATION	00001
QUALITATIVE	00002
QUANTITATIVE	00006
QUESTION	00066
QUESTION NEGOT.	00005
QUESTION-ANSWER	00029
RANDOM	00013
RANDOM-ACCESS	00006
RANK	00017
RATE	00004
READING	00003
REAL-TIME	00001
RECALL	00040
RECOGNITION	00008
RECORD	00008
RECURSIVE	00005
REDUNDANCY	00007
REFERENCE	00027
REJECTION	00006
RELATED	00002

FILMED FROM BEST AVAILABLE COPY

RELATIONSHIP	00040
RELATIVE	00007
RELEV. JUDGEMENT	00004
RELEVANCE	00074
RELEVANT	00031
REMOTE TERMINAL	00010
RESEARCH	00023
RESPONSE TIME	00002
RETRIEVAL	00084
RETRIEVAL SYSTEM	00065
REVIEW	00006
ROLE	00014
ROUTINE	00008
RULE	00013
SAMPLE	00005
SCANNING	00014
SCIENTIFIC	00024
SCOPE NOTE	00002
SEARCH CRITERIA	00002
SEARCH STRATEGY	00035
SEARCHING	00093
SECRECY	00000
SEE ALSO	00002
SEE-REFERENCE	00001
SELECTION	00005
SELECTIVE DISSEM	00006
SEMANTIC	00062
SEQUENCE	00016
SERVICE	00014
SET THEORY	00003
SETS	00016
SHELFLIST	00001
SIGNIFICANCE	00007
SIMULATION	00006
SIZE	00005
SMALL	00001
SMART SYSTEM	00007
SOCIAL IMPLIC.	00002
SOFTWARE	00006
SORTING	00005
SOURCE	00006
SPECIALIZED	00005
SPECIFICITY	00013
STANDARDIZATION	00002
STAT ASSOCIATION	00010
STAT. ANALYSIS	00002
STAT. METHOD	00026
STATE-OF-THE-ART	00011
STATISTICAL	00047
STORAGE	00072
STRING	00012
STRUCTURE	00070
SUBJECT	00022
SUBJECT HEADING	00014
SUBJECT INDEXING	00013
SUBJECT-CATALOG.	00004
SUBJECTIVE	00001

FILMED FROM BEST AVAILABLE COPY

SUBROUTINE	00003
SUMMARY	00002
SURVEY	00012
SYMBOL	00025
SYMBOLIC LOGIC	00004
SYMPOSIUM	00005
SYNONYM	00024
SYNTACTIC ANAL.	00025
SYNTAX	00038
SYSTEM	00102
TABLE	00006
TAG	00016
TECHNICAL	00022
TECHNICAL REPORT	00001
TECHNOLOGY	00012
TELEGRAPHIC ABS.	00004
TERM	00054
TERMINOLOGY	00004
TEST	00025
TEXT	00034
THEORY	00029
THESAURUS	00043
TIME	00009
TIME-SHARING	00010
TITLE	00018
TRANSFORMATION	00025
TRANSLATION	00033
TRANSLITERATION	00000
TRANSMISSION	00004
TREE	00019
TREE STRUCTURE	00014
TRUNCATION	00003
TYPE STYLE	00000
TYPE-SETTING	00003
TYPOGRAPHICAL	00000
UNION CATALOG	00000
UNITERM SYSTEM	00009
UPDATING	00001
USER	00052
USER STUDY	00002
UTILITY	00007
VALIDATION	00002
VALUE	00013
VARIABLE	00016
VECTOR	00015
VENN DIAGRAM	00003
VISUAL DIS. CON.	00000
VOCABULARY	00033
WEIGHT	00031
WEIGHT INDEXING	00007
WORD	00041
WORD ASSOCIATION	00040
WORD FREQUENCY	00012

/*

FILMED FROM BEST AVAILABLE COPY

APPENDIX 3

INDEX TERM FREQUENCY LIST
SORTED ON FREQUENCY OF REFERENCE

FILMED FROM BEST AVAILABLE COPY

INDEX TERM LIST SORTED ON FREQUENCY OF REFERENCE. REVISED APRIL 24, 19

INDEX TERM	NO. OF REFS.
DOCUMENT	00130
COMPUTER	00112
INFO. RETRIEVAL	00105
SYSTEM	00102
INDEXING	00097
SEARCHING	00093
RETRIEVAL	00084
INFORMATION	00083
ANALYSIS	00076
RELEVANCE	00074
EVALUATION	00072
STORAGE	00072
STRUCTURE	00070
CLASSIFICATION	00068
QUESTION	00066
RETRIEVAL SYSTEM	00065
SEMANTIC	00062
LANGUAGE	00059
NATURAL LANGUAGE	00059
AUTOMATIC	00055
INDEX	00054
MATRIX	00054
TERM	00054
EXPERIMENT	00053
USER	00052
ASSOCIATION	00050
DESCRIPTOR	00050
STATISTICAL	00047
DICTIONARY	00044
MEASURE	00043
THESAURUS	00043
PROCESSING	00042
PROGRAM	00041
WORD	00041
RECALL	00040
RELATIONSHIP	00040
WORD ASSOCIATION	00040
HIERARCHY	00039
PROBABILITY	00039
FREQUENCY	00038
MODEL	00038
SYNTAX	00038
INPUT	00037
COMMUNICATION	00036
LIBRARY	00036
LINGUISTIC	00036
ALGORITHM	00035
SEARCH STRATEGY	00035
DATA	00034
TEXT	00034
FILE	00033
TRANSLATION	00033

FILMED FROM BEST AVAILABLE COPY

VOCABULARY	00033
COORDINATE INDEX	00032
GRAMMAR	00031
OUTPUT	00031
RELEVANT	00031
WEIGHT	00031
MATCH	00030
EFFICIENCY	00029
QUESTION-ANSWER	00029
THEORY	00029
AUTO. INDEXING	00028
CONCEPT	00028
PERFORMANCE	00028
PRECISION	00028
ABSTRACT	00027
COMPARISON	00027
REFERENCE	00027
CO-OCCURRENCE	00026
LITERATURE	00026
LOGIC	00026
MATHEMATICAL	00026
STAT. METHOD	00026
SYMBOL	00025
SYNTACTIC ANAL.	00025
TEST	00025
TRANSFORMATION	00025
CONTENT	00024
GRAPH	00024
SCIENTIFIC	00024
SYNONYM	00024
BIBLIOGRAPHIC	00023
KEYWORD	00023
RESEARCH	00023
CLUMP	00022
COEFFICIENT	00022
COST	00022
DESIGN	00022
SUBJECT	00022
TECHNICAL	00022
CLASSIF. SCHEME	00021
CORRELATION	00021
LINK	00021
BOOLEAN	00020
CLUSTER	00020
CODING	00020
HARDWARE	00020
MEANING	00020
EDUCATION	00019
METHODOLOGY	00019
TREE	00019
CODE	00018
LATTICE	00018
LIST	00018
TITLE	00018
ASSOCIATIVE	00017
MAN-MACHINE	00017
MECHANIZATION	00017

FILMED FROM BEST AVAILABLE COPY

RANK	00017
ACCESS	00016
CATEGORIES	00016
CONTEXT	00016
NETWORK	00016
SEQUENCE	00016
SETS	00016
TAG	00016
VARIABLE	00016
BIBLIOGRAPHY	00015
COLLECTION	00015
CONTENT ANALYSIS	00015
EFFECTIVENESS	00015
FUNCTION	00015
INFO. SCIENCE	00015
NEEDS	00015
NOTATION	00015
VECTOR	00015
ALGEBRA	00014
AUTO ABSTRACTING	00014
ERROR	00014
FEEDBACK	00014
FILE ORGANIZATIO	00014
LIBRARIAN	00014
PARSE	00014
ROLE	00014
SCANNING	00014
SERVICE	00014
SUBJECT HEADING	00014
TREE STRUCTURE	00014
ABSTRACTING	00013
CITATION	00013
CITATION INDEX	00013
DOCUMENTATION	00013
GENERIC	00013
INFERENCE	00013
KWIC	00013
ORDER	00013
PROCEDURE	00013
PROG. LANGUAGE	00013
RANDOM	00013
RULE	00013
SPECIFICITY	00013
SUBJECT INDEXING	00013
VALUE	00013
CRITERIA	00012
DISSEMINATION	00012
MECHANICAL	00012
ORGANIZATION	00012
STRING	00012
SURVEY	00012
TECHNOLOGY	00012
WORD FREQUENCY	00012
CHARACTERISTIC	00011
CURRICULUM	00011
DEPTH OF INDEXIN	00011
FALSE DROP	00011

FILMED FROM BEST AVAILABLE COPY

NOISE	00011
PARAMETER	00011
STATE-OF-THE-ART	00011
ADDRESS	00010
ANSWER	00010
AUTOMATION	00010
CRANFIELD	00010
DEGREE	00010
ILLUSTRATION	00010
LOGICAL	00010
REMOTE TERMINAL	00010
STAT ASSOCIATION	00010
TIME-SHARING	00010
BOOK	00009
CARD	00009
CATALOG	00009
COORDINATE	00009
CROSS REFERENCE	00009
CURRENT AWARENES	00009
EDITING	00009
GENERAL	00009
MANUAL	00009
MATHEMATICS	00009
TIME	00009
UNITERM SYSTEM	00009
BINARY	00008
CONNECTION	00008
COUPLING	00008
FACETED CLASSIF.	00008
INTERPRET	00008
INTRODUCTORY	00008
JOURNAL	00008
PERMUTED	00008
PROCEEDINGS	00008
PROFILE	00008
RECOGNITION	00008
RECORD	00008
ROUTINE	00008
CATALOGING	00007
CENTERS	00007
CONDITIONAL PROB	00007
DESCRIPTIVE	00007
ENTRY	00007
FACET	00007
FLOW OF INFO.	00007
INTERDISCIPLINAR	00007
ON-LINE	00007
PHRASE	00007
REDUNDANCY	00007
RELATIVE	00007
SIGNIFICANCE	00007
SMART SYSTEM	00007
UTILITY	00007
WEIGHT INDEXING	00007
ALPHABETIC	00006
AMBIGUITY	00006
ARTIFICIAL LANG.	00006

FILMED FROM BEST AVAILABLE COPY

AUTHOR	00006
COMP LINGUISTICS	00006
CONFERENCE	00006
EXTRACT	00006
FORMAT	00006
IRRELEVANT	00006
LAW	00006
MACHINE-READABLE	00006
NATIONAL	00006
NUMERIC	00006
PATENT	00006
PATTERN	00006
PERTINENT	00006
PREDICTION	00006
QUANTITATIVE	00006
RANDOM-ACCESS	00006
REVIEW	00006
SELECTIVE DISSEM	00006
SIMULATION	00006
SOFTWARE	00006
SOURCE	00006
TABLE	00006
ACCURACY	00005
ACQUISITION	00005
CANONICAL	00005
CHEMICAL	00005
CONTROL	00005
CONTROLLED	00005
DECISION THEORY	00005
DISCRIMINANT	00005
FACT RETRIEVAL	00005
HISTORICAL	00005
INTERACTION	00005
MEETING	00005
OPERATION	00005
PUNCHED	00005
QUESTION NEGOT.	00005
RECURSIVE	00005
SAMPLE	00005
SELECTION	00005
SIZE	00005
SORTING	00005
SPECIALIZED	00005
SYMPOSIUM	00005
ACCESSION NUMBER	00004
APPLICATION	00004
ARRAY	00004
BROWSING	00004
CIRCULATION	00004
CLERICAL	00004
COMBINATIONS	00004
DEDUCTIVE	00004
ENTROPY	00004
GENERATION	00004
INTELLECTUAL	00004
ITERATIVE	00004
MEDIUM	00004

FILMED FROM BEST AVAILABLE COPY

MICROFICHE	00004
OBJECTIVE	00004
PROGRAMMED	00004
RATE	00004
RELEV. JUDGEMENT	00004
SUBJECT-CATALOG.	00004
SYMBOLIC LOGIC	00004
TELEGRAPHIC ABS.	00004
TERMINOLOGY	00004
TRANSMISSION	00004
AUTHORITY LIST	00003
CENTRALIZED	00003
CHANNEL	00003
COMPONENT	00003
CONCORDANCE	00003
CONVENTIONAL	00003
CONVERSION	00003
COUNT	00003
INTERFACE	00003
KEYPUNCH	00003
LABORATORY	00003
MESSAGE	00003
MICROFILM	00003
MORPHOLOGY	00003
MULTIPLE	00003
NON-RELEVANT	00003
OCCURRENCE	00003
OPTIMIZATION	00003
PEEK-A-BOO	00003
PHILOSOPHY	00003
PLANNING	00003
PSYCHOLOGY	00003
READING	00003
SET THEORY	00003
SUBROUTINE	00003
TRUNCATION	00003
TYPE-SETTING	00003
VENN DIAGRAM	00003
ALPHANUMERIC	00002
ANALOGY	00002
ARTIFICIAL INTEL	00002
ASSIGNED	00002
BEHAVIORAL	00002
CARD CATALOG	00002
CYBERNETICS	00002
IDENTIFICATION	00002
INTUITIVE	00002
INVENTORY	00002
INVERTED	00002
LARGE	00002
MANUAL INDEXING	00002
NATURAL	00002
NON-CONVENTIONAL	00002
NUMBER	00002
OFF-LINE	00002
QUALITATIVE	00002
RELATED	00002

FILMED FROM BEST AVAILABLE COPY

RESPONSE TIME	00002
SCOPE NOTE	00002
SEARCH CRITERIA	00002
SEE ALSO	00002
SOCIAL IMPLIC.	00002
STANDARDIZATION	00002
STAT. ANALYSIS	00002
SUMMARY	00002
USER STUDY	00002
VALIDATION	00002
ACTIVITY	00001
ADMINISTRATION	00001
ALPHABETIC ORDER	00001
BATCH PROCESSING	00001
CALL NUMBER	00001
CRITICAL	00001
DECENTRALIZATION	00001
DUAL DICTIONARY	00001
GOVERNMENT	00001
GRAPHICS	00001
INDEPENDENT	00001
LINEAR	00001
MODIFICATION	00001
NON-DISCRIMINANT	00001
NON-FILE	00001
NON-RANDOM	00001
PRINTING	00001
PUNCTUATION	00001
REAL-TIME	00001
SEE-REFERENCE	00001
SHELFLIST	00001
SMALL	00001
SUBJECTIVE	00001
TECHNICAL REPORT	00001
UPDATING	00001
ABBREVIATION	00000
ALTERNATIVES	00000
ANTHOLOGY	00000
CLAIM	00000
COLLOQUIUM	00000
COPYRIGHT	00000
DISSERTATION	00000
IDENTICAL	00000
IMPLEMENTATION	00000
NORMALIZED	00000
PRINCIPLE	00000
REJECTION	00000
SECRECY	00000
TRANSLITERATION	00000
TYPE STYLE	00000
TYPOGRAPHICAL	00000
UNION CATALOG	00000
VISUAL DIS. CON.	00000

FILMED FROM BEST AVAILABLE COPY

APPENDIX 4

A CLASSIFICATION OF THE SUBJECT AUTHORITY LIST

FILMED FROM BEST AVAILABLE COPY

EDITORIAL NOTE

This list has been prepared post hoc from the Subject Authority List as a convenience for students in formulating request inputs for LABSRC3C. When choosing operands for Boolean expressions, it is helpful to have some suggestions for alternative terms, particularly in exercises to observe the effects of variant request specifications on the output. The groupings in this list are offered as a way of reducing the need to read through the entire Subject Authority List in order to obtain hints for editing requests. Their purpose is utilitarian, and the list should not be interpreted as representing any carefully formulated judgment concerning the synonymity or hierarchical relationships of the terms.

To get the most benefit from this list, it should be used in connection with the list of term frequencies in Appendix 2.

FILMED FROM BEST AVAILABLE COPY

I. INFORMATION, COMMUNICATION

1. general interdisciplinary
   philosophy theory
2. administration organization control system
   government national law
   centers centralized network
   service library retrieval system
3. context social impl. historical
4. education curriculum research
5. flow of info. dissemination circulation access
   source author generation
   reference citation title
   bibliography documentation
   selection acquisition
   selective dissem. current awareness
6. literature document book journal
   technical specialized scientific chemical
   technical specialized scientific chemical
   technical report patent
   microfilm microfiche
   printing editing graphics
7. message data content
   transmission channel medium
   feedback modification
   redundancy entropy
   noise error ambiguity

FILMED FROM BEST AVAILABLE COPY

II. PSYCHOLOGY

1. behavioral activity
2. intellectual scientific logical
   subjective qualitative intuitive
3. user user study
   characteristic profile
   relevance judgment criteria
4. needs utility value

III. ACCESS, RETRIEVAL

1. fact retrieval info. retrieval
   cybernetics artificial intelligence question-answer
   question search criteria input
   recognition comparison match
   answer feedback output
   (see also VII.5
2. searching search strategy
   question negot. modification
   iterative recursive
   browsing scanning reading
3. relevance
   selection rank order
   relevant pertinent
   irrelevant non-relevant false drop
   association relationship connection
4. retrieval system
   a) design planning optimization
      variable component parameter

FILMED FROM BEST AVAILABLE COPY

b) performance

objective function

value utility cost

time response-time rate

c) evaluation measure analysis validation

methodology procedure

test experiment research comparison

d) performance

effectiveness efficiency

accuracy precision recall

e) conventional clerical manual

f) non-conventional automatic

Cranfield Smart System

(see also VII.1)

5. File Organization (see also VII.4)

a) file catalog

size large small

b) sequence order format alphabetic order

linear inverted non-random peek-a-boo

sorting alphabetic alphanumeric numeric

c) entry access address

record card

IV. LANGUAGE, LINGUISTIC (see also V.1)

1. natural language text

word symbol string context

FILMED FROM BEST AVAILABLE COPY

2. comp linguistics word frequency occurrence
   word association co-occurrence
3. grammar structure
   transformation morphology
   syntax syntactic anal. parse pattern
   punctuation notation
4. semantic meaning concept
   content analysis interpret analogy
   ambiguity translation

V. CONTENT ANALYSIS, INDEXING, ABSTRACTING

1. vocabulary terminology
   dictionary thesaurus authority list
   canonical controlled
   entry descriptor subject heading
   term tag keyword
2. cross reference see reference see also synonym
   related relationship relative
   scope note discriminant role link
   (see also IV, V.5)
3. cataloging
   recognition identification descriptive
   subject-catalog. subject indexing
   depth of indexing specificity significance
   weight indexing auto. indexing
   manual indexing card catalog

FILMED FROM BEST AVAILABLE COPY

4. index concordance
   permuted kwic
   coordinate index uniterm system dual dictionary
   shelflist inventory
   catalog list bibliography bibliographic
   citation index coupling (see also I.5)
5. abstract extract abstracting
   auto abstracting telegraphic abs.
6. classification (see also VI.3)
   classif. scheme faceted classif.
   lattice hierarchy tree structure
   generic relative
   clump cluster
   categories subject facet characteristic
   notation call number accession number

VI. MATHEMATICS

1. logic logical mathematical
   deductive inference interpret
   sets set theory Venn diagram
   symbolic logic Boolean
2. quantitative measure count
   number numeric binary
   coefficient weight degree
3. algebra
   variable parameter
   lattice matrix table array tree
   graph vector connection

FILMED FROM BEST AVAILABLE COPY

4. probability prediction

independent random occurrence frequency

conditional prob. decision theory

statistical stat. method stat. analysis sample

stat. association co-occurrence correlation

combinations pattern non-random

association associative related

5. model simulation

VII. COMPUTER

1. hardware technology

automation mechanization

automatic programmed mechanical

2. software

program routine subroutine

algorithm rule

prog. language artificial lang. code

3. input off-line

coding conversion editing truncation machine-readable

keypunch card punched

4. processing

order sequence array sorting

storage address random-access

batch-processing updating real-time

5. output

on-line time-sharing interface interaction

man-machine remote terminal

FILMED FROM BEST AVAILABLE COPY

VIII. REJECTS

The following terms show some uncertainty or incompleteness in their application. For the time being, we suggest not searching on them.

assigned critical illustration

conference meeting colloquium symposium proceedings

review survey summary collection introductory state-of-the-art

natural non-file non-discriminant